Stochastic Methods For Optimization and Machine Learning

Zdravko Botev, BSc

7 November 2005

A thesis written during the academic year of March-November 2005 and submitted for the degree of Bachelor of Science with Honours at the School of Physical Sciences, Department of Mathematics, The University of Queensland, Australia.

Acknowledgments

I would like to thank my parents for their love and support and my supervisor Dr. Dirk Kroese and high-school teacher Ivan Georgiev for their inspirational scholarship, constant encouragement and great tutelage.

Abstract

In this project a stochastic method for general purpose optimization and machine learning is described. The method is derived from basic information-theoretic principles and generalizes the popular Cross Entropy method. The effectiveness of the method as a tool for statistical modeling and Monte Carlo simulation is demonstrated with an application to the problems of density estimation and data modeling.

Keywords

Maximum Entropy, Cross Entropy, measures of information, Monte Carlo simulation, statistical modeling, machine learning, CE method, kernel smoothing, regularization theory, functional optimization

Contents

Acronyms

PWC	Piecewise Continuous
PWS	Piecewise Smooth
CE	Cross Entropy
GCE	Generalized Cross Entropy
MCE	Minimum Cross Entropy
MSE	Mean Squared Error
AMSE	Asymptotic Mean Squared Error
MISE	Mean Integrated Squared Error
AMISE	Asymptotic Mean Integrated Squared Error
QPP	Quadratic Programming Problem
LPP	Linear Programming Problem
GPP	Geometric Programming Problem
MVUE	Minimum Variance Unbiased Estimator/Estimation
pmf	probability mass function
pdf	probability density function
LCP	Linear Complementarity Problem
IS	Importance Sampling
KKT	Karush-Kuhn-Tucker conditions
CMC	Crude Monte Carlo

Notation

$\mathcal{X}_n$	a sample of n random variables or empirical observations
$\mathscr{X}$	set over which the stochastic model works
$\mathbf{x} \in \mathscr{X}$	d dimensional column vector
$\int$	$\int\cdots\int$- or one -dimensional integral (depending on context)
$d\mathbf{x}$	$dx_1\, dx_2 \ldots dx_d$
p	proposal/sampling/instrumental distribution
q	a-prori distribution
q^*	optimal importance sampling distribution
$K_i : \mathscr{X} \to \mathbb{R}^+$	kernel function anchored at the i-th datum
$\mathcal{K}$	univariate kernel function
$\mathbf{K}(\mathbf{x})$	$[K_1(\mathbf{x}), \ldots, K_n(\mathbf{x})]^T$
κ_i	$\kappa_i = \mathbb{E}_q K_i(\mathbf{X})$
$\boldsymbol{\kappa}$	$\boldsymbol{\kappa} = \mathbb{E}_q \mathbf{K}(\mathbf{X})$
κ_i^*	κ_i^* is an estimate of $\mathbb{E}_{q^*} K_i(\mathbf{X})$
$\boldsymbol{\kappa}^*$	$\boldsymbol{\kappa}^*$ is an estimate of $\mathbb{E}_{q^*} \mathbf{K}(\mathbf{X})$
$\boldsymbol{\lambda}$	the set of Lagrange multipliers
$\boldsymbol{\nu}$	the set of rescaled Lagrange multipliers
Σ_i	scale/bandwidth matrix of K_i
$\mathcal{D}$	Csiszár's distance measure
$\mathcal{L}$	Lagrangian of the primal optimization problem
$\mathcal{L}^*(\boldsymbol{\lambda}, \lambda_0)$	$\mathcal{L}^*(\boldsymbol{\lambda}, \lambda_0) = \inf_p \mathcal{L}(p; \boldsymbol{\lambda}, \lambda_0)$
C^n	set of n times differentiable functions
$\mathsf{N}(\boldsymbol{\mu}, \Sigma)$	multivariate normal with mean $\boldsymbol{\mu}$ and covariance Σ
C	covariance matrix for the QPP
$\mathbf{c}$	$\mathbf{c} = \boldsymbol{\kappa}^* - \boldsymbol{\kappa}$
V	$V = \mathrm{diag}(C)$
A	correlation matrix corresponding to C
$\mathbf{a}$	$\mathbf{a} = V^{-1/2}\mathbf{c}$
$\mathscr{P}$	The set of valid probability densities on $\mathscr{X}$
$\mathscr{S}$	the set of admissible bandwidth parameters

Introduction

In a series of papers and books (see [34], [36], [37], [35], [31], [29], [27], [57], [38], [30], [33], [28], [32]), the most notable of which are [38], [32] and [30], Kapur and Kesavan described a generalization of the Maximum Entropy Method of Jaynes [26] and the information-theoretic concepts of Shannon [58].

The main goal of this project is to describe a stochastic optimization and machine learning method which fuses these generalized information-theoretic concepts with traditional Monte Carlo simulation. The method is fundamentally a generalization of the *Cross Entropy* (CE) method [54]. Due to its information-theoretic derivation and generalization property, the method will be referred to as the *Generalized Cross Entropy* method (GCE). The GCE is designed to provide a solution to the following problems in a simple unified framework:

Monte Carlo Simulation The major problem is to sample from an arbitrary probability function q^* given that we can evaluate q^* up to an unknown constant. The most efficient method for sampling from such a q^* will also be the most efficient and generally applicable method for stochastic simulation.

Statistical Learning The main problem is to find/estimate the sparsest probability model q^* for a given empirical data with as little loss of information as possible. This problem is usually harder than the Monte Carlo Simulation problem because the 'true' distribution of the empirical data is unknown.

Example 1 (Integration in multiple dimensions) The problem is to estimate integrals of the form:

$$\int_{\mathscr{X}} H(\mathbf{x})\, d\mathbf{x},$$

for an arbitrary function H. Alternatively the discrete analogue is to estimate

$$\sum_{\mathbf{x}\in\mathscr{X}} H(\mathbf{x}), \tag{1}$$

where the set $\mathscr{X}$ can be so large that straightforward summation is impractical. E.g., estimation of (1) for $H(\mathbf{x}) = I\{\mathbf{x} \in \mathscr{X}^*\}$, where $\mathscr{X}^* \subset \mathscr{X}$, is a class of important and difficult discrete counting problems. These problems can be efficiently solved by (approximately) sampling from $q^*(\mathbf{x}) = c\,|H(\mathbf{x})|$, where c is an unknown normalizing constant.

Example 2 (Rare-event Simulation) Rare event simulation is a special case of integration in multiple dimensions:

$$\ell = \int_{\mathscr{X}} \varphi(S(\mathbf{x});\gamma) f(\mathbf{x})\, d\mathbf{x} = \int_{\mathscr{X}} H(\mathbf{x}) f(\mathbf{x})\, d\mathbf{x} = \mathbb{E}_f H(\mathbf{X}),$$

where ℓ is small and f is a light- or heavy-tailed probability density, φ a real-valued function depending on a real-valued function S and a parameter. Again this problem is solved efficiently by sampling from the *minimum variance* importance sampling density [54]:

$$q^*(\mathbf{x}) = c\,|H(\mathbf{x})|\,f(\mathbf{x})\ .$$

Example 3 (Optimization) Global optimization of non-smooth or discrete multidimensional multimodal functions. For example, the GCE may help simulate variates from the density

$$q^*(\mathbf{x}) = \frac{I_{\{S(\mathbf{x})>\gamma\}}}{\sum_{\mathscr{X}} I_{\{S(\mathbf{x})>\gamma\}}}\ .$$

This equates to knowledge of the set over which a given function $S : \mathscr{X} \to \mathbb{R}$ takes values above γ.

Example 4 (Random Variate Generation) Efficient generation of random variables from complicated continuous or discrete probability functions via the Accept–Reject method. For example simulate random variables from the Boltzmann-Shannon density:

$$q^*(\mathbf{x}) = \frac{\mathrm{e}^{-\lambda S(\mathbf{x})}}{\int_{\mathscr{X}} \mathrm{e}^{-\lambda S(\mathbf{x})}\,d\mathbf{x}},$$

where $\lambda \in \mathbb{R}$ is the annealing constant and $S : \mathscr{X} \to \mathbb{R}$.

Example 5 (Statistical modeling) The problem of statistical data analysis, which can be stated as follows: Given a finite number of empirical observations $\mathcal{X}_n = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$:

1. either find a few elements in the set $\mathcal{X}_n$ which are representative of the whole set, i.e., classify and/or compress the data,

2. or find the optimal (in some sense) probability model for the data, i.e., estimate the probability function for which the data is assumed to be a random outcome. Once the probability function is estimated non-parametric inference (hypothesis testing, confidence bands etc.) using the theory of smoothed bootstrap [19] can be conducted.

Both Statistical Learning and Monte Carlo Simulation can be *ill-posed* problems in the sense that extra assumptions need to be introduced for unique stable and well-behaved solutions to exist. Some possible approaches to ill-posed problems are:

1. Regularization theory as described by Vapnik [66].

2. The array of information-theoretic methods described in [33].

In this project we will only consider the information-theoretic approach. To this end we first review some relevant background material.

1 Preliminaries

1.1 Stratified Sampling

Stratified sampling is a method of reducing the variance of statistical estimators. In the derivation of the method, the following result is used.

Lemma 1 (Conditional Variance) For any random variables $\mathbf{X}$ and $\mathbf{Y} \in \mathscr{X}$:

$$\mathrm{Var}(H(\mathbf{X})) = \mathbb{E}[\mathrm{Var}(H(\mathbf{X}) \,|\, \mathbf{Y})] + \mathrm{Var}(\mathbb{E}[H(\mathbf{X}) \,|\, \mathbf{Y}]),$$

where $H : \mathscr{X} \to \mathbb{R}$. Hence $\mathrm{Var}(H(\mathbf{X})) \geqslant \mathrm{Var}(\mathbb{E}[H(\mathbf{X}) \,|\, \mathbf{Y}])$ and $\mathrm{Var}(H(\mathbf{X})) \geqslant \mathbb{E}[\mathrm{Var}(H(\mathbf{X}) \,|\, \mathbf{Y})]$.

Suppose we want to estimate

$$\ell = \int_{\mathscr{X}} H(\mathbf{x})\, p(\mathbf{x})\, d\mathbf{x} = \mathbb{E}_p[H(\mathbf{X})]$$

and that p can be written in the form $p(\mathbf{x}) = \sum_{k=0}^{n} p(\mathbf{x}, k) = \sum_{k=0}^{n} p(\mathbf{x} \,|\, k)\, p(k)$. Then by the tower property

$$\ell = \mathbb{E}\left[\mathbb{E}[H(\mathbf{X}) \,|\, \mathrm{K}]\right] = \sum_{k=0}^{n} p(k)\, \mathbb{E}[H(\mathbf{X}) \,|\, \mathrm{K} = k].$$

This suggests that, using a fixed budget of N samples, we can estimate ℓ

1. either using the *Crude Monte Carlo* (CMC) estimator:

$$\hat{\ell} = \frac{1}{N} \sum_{i=1}^{N} H(\mathbf{X}_i), \quad \mathbf{X}_1, \ldots, \mathbf{X}_N \overset{i.i.d}{\sim} p,$$

2. or the *stratified* estimator:

$$\hat{\ell}_s = \sum_{k=0}^{n} p(k)\, \frac{1}{N_k} \sum_{j=1}^{N_k} H(\mathbf{X}_{kj}),$$

where $\{\mathbf{X}_{kj}\}_{j=1}^{N_k} \overset{i.i.d}{\sim} p(\mathbf{x} \,|\, k)$ for each k and $\sum_{k=0}^{n} N_k = N$.

Depending on the choice of $\{N_k\}_{k=1}^{n}$, the stratified estimator can perform better than the CMC estimator. To see this note that:

$$\mathrm{Var}(\hat{\ell}_s) = \sum_{k=0}^{n} \frac{p^2(k)}{N_k}\, \mathrm{Var}(H(\mathbf{X}) \,|\, \mathrm{K} = k) = \sum_{k=0}^{n} \frac{p^2(k)}{N_k}\, \sigma_k^2.$$

We now wish to choose the set $\{N_k\}_{k=0}^n$ such that the variance of $\hat{\ell}_s$ is as small as possible subject to the budget constraint that $\sum_{k=0}^n N_k = N$. The Lagrange multiplier technique gives the (approximate) optimal solution

$$N_k^* = N \times \frac{p(k)\ \sigma_k}{\sum_{k=0}^n p(k)\ \sigma_k}$$

with corresponding minimal variance

$$\min_{N_1,\ldots,N_k} \mathrm{Var}(\hat{\ell}_s) = \frac{1}{N}\left[\sum_{k=0}^n p(k)\ \sigma_k\right]^2 = \frac{1}{N}\Big[\mathbb{E}[\sigma_{\mathrm{K}}]\Big]^2.$$

In the special case where $\sigma_k = \sigma, \forall k$, the optimal $N_k^* = N \times p(k)$ with corresponding variance $\frac{\sigma^2}{N}$. With slight abuse of the $\mathrm{Var}(\cdot\,|\,\cdot)$ notation:

$$\begin{aligned}
N \times \mathrm{Var}(\hat{\ell}) &= \mathrm{Var}(H(\mathbf{X})) && (2)\\
&\geqslant \mathbb{E}[\mathrm{Var}(H(\mathbf{X})\,|\,K)] \qquad \text{by lemma 1} && (3)\\
&= \sum_{k=0}^n p(k)\ \sigma_k^2 = \mathbb{E}\left[\sigma_{\mathrm{K}}^2\right] = N \times \mathrm{Var}\left(\hat{\ell}_s\,|\,N_k \propto N \times p(k)\right) && (4)\\
&\geqslant \Big[\mathbb{E}[\sigma_{\mathrm{K}}]\Big]^2 = N \times \mathrm{Var}\left(\hat{\ell}_s\,|\,N_k \propto N \times p(k) \times \sigma_k\right). && (5)
\end{aligned}$$

Hence we have the relation:

$$\mathrm{Var}\left(\hat{\ell}\right) \geqslant \mathrm{Var}\left(\hat{\ell}_s\,|\,N_k \propto N \times p(k)\right) \geqslant \mathrm{Var}\left(\hat{\ell}_s\,|\,N_k \propto N \times p(k) \times \sigma_k\right).$$

Thus stratification will always improve on the CMC estimation of ℓ. In practice neither $\{\omega_k = N \times p(k)\}_{k=1}^n$ nor $\left\{\omega_k = N \times \frac{p(k)\times\sigma_k}{\sum_{k=0}^n p(k)\ \sigma_k}\right\}_{k=1}^n$ are integers. Instead of rounding $\{\omega_k\}_{k=1}^n$ we can allocate random $\{N_k\}_{k=1}^n$ such that $\mathbb{E}[N_k] = \omega_k$ using the following algorithm:

Algorithm 1.1 (Stratified Sampling)

1. Generate $\lfloor\omega_k\rfloor$ random variables from each $p(\mathbf{x}\,|\,k)$ to obtain a total of $\sum_{i=1}^n \lfloor\omega_k\rfloor$ random variables.
2. We can generate $N - \sum_{i=1}^n \lfloor\omega_k\rfloor$ more random variables before exhausting the budget. The residual number of random variables $r = N - \sum_{i=1}^n \lfloor\omega_k\rfloor$ is obtained in the following way. Sample r indexes $\{\mathrm{K}_i\}_{i=1}^r$ with replacement from the set $\{1,\ldots,n\}$ with probabilities proportional to $\{\omega_k - \lfloor\omega_k\rfloor\}_{k=1}^n$. Using the random set of indexes $\{\mathrm{K}_i\}_{i=1}^r$ generate r more random variables from the set $\{p(\mathbf{x}\,|\,k)\}_{k=1}^n$.

This is the algorithm which we intend to use in order to reduce sampling variability. Stratification is the incarnation of the common sense idea that if we have to integrate a complicated integrand then we should try to do as much of the integration deterministically and use Monte Carlo only for the parts of the integrand which do not yield to deterministic quadrature methods.

1.2 Properties of Multivariate Gaussian Density

Here we derive some simple, yet little used and known, properties of the multivariate Gaussian density. We use these results in subsequent sections. Let $\phi(\mathbf{x}) = (2\pi)^{-d/2} \exp\left(-\frac{1}{2}\mathbf{x}'\mathbf{x}\right)$, where $\mathbf{x} \in \mathbb{R}^d$ is a column vector, denote the multivariate $\mathsf{N}(\mathbf{0}, \mathbf{I})$ density. Then $|\Sigma|^{-1/2}\phi\left(\Sigma^{-1/2}\left(\mathbf{x} - \boldsymbol{\mu}\right)\right) \equiv \mathsf{N}(\boldsymbol{\mu}, \Sigma)$ and

$$\begin{aligned}
&\int |\Sigma_1|^{-1/2}\phi\left(\Sigma_1^{-1/2}\left(\mathbf{x} - \boldsymbol{\mu}_1\right)\right) \times |\Sigma_2|^{-1/2}\phi\left(\Sigma_2^{-1/2}\left(\mathbf{x} - \boldsymbol{\mu}_2\right)\right) d\mathbf{x} \\
&= (2\pi)^{-d}|\Sigma_1\Sigma_2|^{-1/2} \times \\
&\int \exp\left(\mathbf{x}'\left(\Sigma_1^{-1} + \Sigma_2^{-1}\right)\mathbf{x} - 2\mathbf{x}'\left(\Sigma_1^{-1}\boldsymbol{\mu}_1 + \Sigma_2^{-1}\boldsymbol{\mu}_2\right) + \boldsymbol{\mu}_1'\Sigma_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2'\Sigma_2^{-1}\boldsymbol{\mu}_2\right)^{-1/2} d\mathbf{x} \\
&= (2\pi)^{-d}|\Sigma_1\Sigma_2|^{-1/2} \times \\
&\int \exp\left(\mathbf{x}'\Sigma^{-1}\mathbf{x} - 2\mathbf{x}'\Sigma^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu} - \boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}_1'\Sigma_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2'\Sigma_2^{-1}\boldsymbol{\mu}_2\right)^{-1/2} d\mathbf{x} \\
&= (2\pi)^{-d}|\Sigma_1\Sigma_2|^{-1/2} \times (2\pi)^{d/2}|\Sigma|^{1/2} \exp\left(\boldsymbol{\mu}_1'\Sigma_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2'\Sigma_2^{-1}\boldsymbol{\mu}_2 - \boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu}\right)^{-1/2},
\end{aligned}$$

where $\Sigma^{-1} = \Sigma_1^{-1} + \Sigma_2^{-1}$ and $\boldsymbol{\mu} = \Sigma\left(\Sigma_1^{-1}\boldsymbol{\mu}_1 + \Sigma_2^{-1}\boldsymbol{\mu}_2\right)$. Thus in general the integral of the product of n multivariate Gaussian densities gives:

$$\int \prod_{i=1}^{n} \mathsf{N}(\boldsymbol{\mu}_i; \Sigma_i)\, d\mathbf{x} = \frac{(2\pi)^{\frac{d(1-n)}{2}}\,|\Sigma|^{\frac{1}{2}}}{\prod_{i=1}^{n}|\Sigma_i|^{\frac{1}{2}}} \exp\left(\frac{1}{2}\boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu} - \frac{1}{2}\sum_{i=1}^{n}\boldsymbol{\mu}_i'\Sigma_i^{-1}\boldsymbol{\mu}_i\right), \tag{6}$$

where $\Sigma^{-1} = \sum_{i=1}^{n}\Sigma_i^{-1}$ and $\boldsymbol{\mu} = \Sigma\sum_{i=1}^{n}\Sigma_i^{-1}\boldsymbol{\mu}_i$. After some tedious algebra it can be shown that for $n = 2$:

$$\int \prod_{i=1}^{2} \mathsf{N}(\boldsymbol{\mu}_i; \Sigma_i)\, d\mathbf{x} = \mathsf{N}\left(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2; \Sigma_1 + \Sigma_2\right). \tag{7}$$

The product of n Gaussian densities is proportional to another Gaussian:

$$\prod_{i=1}^{n} \mathsf{N}(\boldsymbol{\mu}_i; \Sigma_i) = c \times \mathsf{N}(\boldsymbol{\mu}; \Sigma), \tag{8}$$

where $c = \frac{(2\pi)^{\frac{d(1-n)}{2}}|\Sigma|^{\frac{1}{2}}}{\prod_{i=1}^{n}|\Sigma_i|^{\frac{1}{2}}} \exp\left(\frac{1}{2}\boldsymbol{\mu}'\Sigma^{-1}\boldsymbol{\mu} - \frac{1}{2}\sum_{i=1}^{n}\boldsymbol{\mu}_i'\Sigma_i^{-1}\boldsymbol{\mu}_i\right)$.

1.3 The Euler-Lagrange Equation

Here we review some basic results from the theory of Calculus of Variations as described in [50] and [67]. We start by stating the basic Calculus of Variations problem in the one dimensional case.

Definition 1 (Basic Problem) Find a function $y(t)$ from a specified set of *comparison functions* on the interval $[t_0, t_1]$ which minimizes the integral

$$J[y] = \int_{t_0}^{t_1} L\left(y(t), \dot{y}(t), t\right)\ dt.$$

In other words the problem is:

$$\min_{y \in \mathscr{Y}} J[y],$$

where $\mathscr{Y}$ is the set of *admissible comparison functions*. In many practical situations $y(t)$ has to satisfy the boundary conditions $y(t_0) = y_0,\ y(t_1) = y_1$ for some fixed y_0 and y_1.

For different sets $\mathscr{Y}$, the basic problem usually has different solutions. For our purposes we focus attention on the set of piecewise smooth functions:

Definition 2 (Piecewise Smooth Function) A function $y(t)$ is said to be piecewise smooth on the interval $[a, b]$ if :

1. It is continuous on $[a, b]$.
2. The derivative $\dot{y}$ fails to exist at at most a finite number of points in $[a, b]$. I.e., $\dot{y}$ is *piecewise continuous* (PWC)— continuous over a finite number of subintervals.

A necessary condition for a solution of the basic problem in the class of PWS functions is the Euler -Lagrange equation:

Theorem 1 (Euler-Lagrange) In order that $y^*(t)$ minimizes the functional $J[y] = \int_{t_0}^{t_1} L(y(t), \dot{y}(t), t)\ dt$ in the class of piecewise continuous functions, it is necessary that the Euler-Lagrange equation:

$$\frac{\partial L}{\partial y} - \frac{d}{dt}\left(\frac{\partial L}{\partial \dot{y}}\right) = 0$$

holds at each point of $y^*(t)$ for which $\dot{y}(t)$ is continuous. Then $y^*(t)$ is called an *extremal* of $J[y]$ in the set of admissible comparison functions $\mathscr{Y}$.

For a proof of this see [48]. We now have the following important theorem concerning sufficient conditions (see [67] page 108):

Theorem 2 (Sufficient Conditions) For differentiable functions $L(y, \dot{y}, t)$ convex in both y and $\dot{y}$, any admissible extremal $y^*(t) \in \mathscr{Y}$ renders $J[y^*]$ a global minimum.

1.3.1 Inequality Constraint

Suppose we now modify the basic problem to include a pointwise inequality constraint:

Definition 3 (Pointwise Inequality Constraint) The basic problem with the addition of a pointwise inequality constraint is:

$$\begin{aligned} &\min_{y \in \mathscr{Y}} J[y] \\ &\quad y(t_0) = y_0 \\ &\quad y(t_1) = y_1 \\ &\quad y(t) \geqslant \varphi(t), \quad \forall t \in [t_0, t_1], \end{aligned}$$

where y_0 or y_1 could be fixed in advance or allowed to take arbitrary values.

This additional constraint usually complicates the basic problem enormously. Some relevant results are :

Theorem 3 (Inequality Constraint I) Let $\check{y}(t) \in \mathscr{Y}$ minimize $J[y]$ subject to the inequality constraint $y(t) \geqslant \varphi(t),\ \forall t \in [t_0, t_1]$, then $\check{y}(t)$ consists of segments of $y^*(t)$ and segments of $\varphi(t)$. At the *switch points* which join the two different types of segments , $\check{y}(t)$ is continuous.

Here again $y^*(t)$ denotes an admissible extremal of $J[y]$. An elaboration of this result is the following theorem:

Theorem 4 (Inequality Constraint II) If $y^*(t)$ violates the inequality constraint $y(t) \geqslant \varphi(t)$ in a set $\left(\bar{c}, \bar{d}\right) \subset [a, b]$, then the true solution $\check{y}(t)$ must be equal to $\varphi(t)$ in $(c, d) \subseteq \left(\bar{c}, \bar{d}\right)$.

We still need to determine the location of the *switch points* c and d at which the pointwise inequality constraint is enforced and we switch from the extremal $y^*(t)$ to $\varphi(t)$.

Theorem 5 (Optimal Switch Point) An optimal switch point $\check{c}$ between $\varphi(t)$ and $y^*(t)$ is determined by:

1. the *transversality condition*

$$L\left(y^*(\check{c}), \dot{y}^*(\check{c}), \check{c}\right) - L\left(\varphi(\check{c}), \varphi(\check{c}), \check{c}\right) - [\dot{y}^*(\check{c}) - \dot{\varphi}(\check{c})] \frac{\partial L}{\partial \dot{y}}\left(y^*(\check{c}), \dot{y}^*(\check{c}), \check{c}\right) = 0$$

2. and the *continuity requirement*

$$y^*(\check{c}) = \varphi(\check{c}).$$

We thus know that the solution of the basic problem with the addition of a pointwise inequality constraint has the form:

$$\breve{y}(t) = \begin{cases} y^*(t), & t \in \mathcal{S} \subset [a,b] \\ \varphi(t), & t \in \{\mathcal{S} \cap [a,b]\}^c \end{cases} ,$$

where the boundary of the set $\mathcal{S}$ is determined by the optimal switching points and $y^*(t) \geqslant \varphi(t),\ \forall t \in \mathcal{S}$.

1.3.2 Extensions to Multiple Dimensions

Suppose that $y : \mathbb{R}^d \to \mathbb{R}$ and that we wish to find a necessary condition for a solution to the problem

$$\min_{y \in \mathscr{Y}} J[y] ,$$

where $J[y] = \int_{\mathscr{X}} L\left(y(\mathbf{x}), \nabla_{\mathbf{x}} y(\mathbf{x}), \mathbf{x}\right) d\mathbf{x}$ with $\mathbf{x} = \sum_{i=1}^{d} x_i \mathrm{e}_i$ and $\nabla_{\mathbf{x}} = \sum_{i=1}^{d} \frac{\partial}{\partial x_i} \mathrm{e}_i$. A necessary condition is given by the analogue of the Euler-Lagrange equation in multiple dimensions (see [67] page 455):

Theorem 6 (Euler-Lagrange in $\mathbb{R}^d$) A necessary condition for a $\mathscr{Y} \equiv C^1$ admissible extremal of $J[y]$ is :

$$\frac{\partial L}{\partial y} - \nabla_{\mathbf{x}} \cdot \nabla_{\dot{\mathbf{y}}} L = 0,$$

where $\dot{\mathbf{y}} = \nabla_{\mathbf{x}} y(\mathbf{x})$.

1.4 The Rayleigh-Ritz Method

The solution of the basic problem in terms of simple known functions is rarely possible. A numerical approximation to the true solution of the basic problem can be obtained in one of the following ways:

1. Numerical solution of the Euler-Lagrange differential equation (usually a Boundary Value Partial Differential Equation).

2. A direct approach in which the integral is discretized using a fine mesh. This approach is only rarely used due to its computational cost.

3. Using Dynamic Programming to solve an associated *shortest route* problem (see [67]). The shortest route problem is a discrete optimization problem.

4. The Rayleigh-Ritz method as described in [67]. The approach is similar to the finite element method for solving Partial Differential Equations. The GCE method resembles the Rayleigh-Ritz method and in its essence is most probably the Rayleigh-Ritz method in disguise. For this reason we briefly describe the Rayleigh-Ritz method.

The idea behind the Rayleigh-Ritz method is to search for a minimizer of $J[y]$ within a convenient space spanned by simple admissible comparison functions. It can be shown that using a judiciously chosen set of simple *coordinate functions* $\{K_k\}_{k=1}^n$ (see [67] page 202),

$$y_n(t) = \sum_{k=1}^{n} \omega_k \, K_k(t)$$

converges to the true solution $y^*(t)$ as $n \to \infty$. We simply have to determine the coefficients $\{\omega_k\}_{k=1}^n$. This is easily done by substituting the approximate solution into J:

$$J[y_n] = \int_{t_0}^{t_1} L(y_n(t), \dot{y}_n(t), t) \, dt = J\left[\{\omega_k\}_{k=1}^n\right].$$

The coordinate functions are usually simple and the integral can be computed analytically giving a function of the unknown coefficients. Then the infinite dimensional Calculus of Variations problem reduces to the finite parameter optimization problem:

$$\min_{\{\omega_k\}_{k=1}^n} \quad J\left[\{\omega_k\}_{k=1}^n\right] \tag{9}$$

$$\text{subject to:} \quad \sum_{k=1}^{n} \omega_k \, K_k(t_0) = y_0 \tag{10}$$

$$\sum_{k=1}^{n} \omega_k \, K_k(t_1) = y_1. \tag{11}$$

Using standard optimization algorithms the problem can be solved to give the approximate solution $y_n(t) = \sum_{k=1}^{n} \omega_k^* \, K_k(t)$.

1.5 Convex Optimization and Duality

Optimization, whether it be subject to constraints or not, is of utmost importance in applied mathematics. The structure of most optimization problems can be summarized as:

Definition 4 (Basic Optimization Problem)

$$\min_{\mathbf{x}\in\mathbb{R}^d} \; f(\mathbf{x}) \tag{12}$$

$$\text{subject to:} \quad c_i(\mathbf{x}) = 0, \; i \in \mathcal{E} \tag{13}$$

$$c_i(\mathbf{x}) \geqslant 0, \; i \in \mathcal{I}. \tag{14}$$

Within this formulation fall many of the traditional optimization problems, the simplest possible of which are:

1. Linear Programming (LP), in this case f and c_i are linear functions. The *standard form* of all Linear Programming problems is:

$$\begin{aligned} \min_{\mathbf{x}} \; & \mathbf{c}^T\mathbf{x} \\ \text{subject to:} \quad A\mathbf{x} &= \mathbf{b} \\ \mathbf{x} &\geqslant \mathbf{0}, \end{aligned}$$

   where A is an $n \times d$ matrix (usually $n < d$) and $\mathbf{c} \in \mathbb{R}^d$ is a column vector of coefficients.

2. Quadratic Programming Problem (QPP), in this case f is a quadratic function and c_i are linear:

$$\begin{aligned} \min_{\mathbf{x}} \; & \frac{1}{2}\mathbf{x}^T C\mathbf{x} + \mathbf{c}^T\mathbf{x} \\ \text{subject to:} \quad \mathbf{a}_i^T\mathbf{x} &= b_i, \quad i \in \mathcal{E} \\ \mathbf{a}_i^T\mathbf{x} &\geqslant b_i, \quad i \in \mathcal{I}, \end{aligned}$$

   Quadratic programming differs from LP in that it is possible to have meaningful problems in which there are no inequality constraints.

3. Geometric Programming Problem (GPP):

$$\min_{\mathbf{x}} \; f(\mathbf{x}) \tag{15}$$

$$\text{subject to:} \quad g_i(\mathbf{x}) = 1, \quad \forall i \tag{16}$$

$$h_j(\mathbf{x}) \leqslant 1, \quad \forall j \tag{17}$$

$$\mathbf{x} > \mathbf{0}, \tag{18}$$

   where f and $\{h_j\}$ are functions of the form

$$\sum_{k=1}^{K} \omega_k \, x_1^{a_{1k}} \cdots x_d^{a_{dk}}, \; \omega_k > 0, \; \{a_{ij}\} \in \mathbb{R}$$

   and $\{g_i\}$ are functions of the form

$$w \, x_1^{b_1} \cdots x_d^{b_d}, \; w > 0, \; \{b_i\} \in \mathbb{R}.$$

In this project we will be mostly concerned with the QPP. The Lagrangian approach to the solution of the basic optimization problem (12) is to define the *Lagrangian* function $\mathcal{L}(\mathbf{x};\boldsymbol{\lambda}) = f(\mathbf{x}) - \sum_i \lambda_i c_i(\mathbf{x})$. Then a necessary condition for a local solution is:

Theorem 7 (Karush-Kuhn-Tucker conditions) Under mild regularity conditions, there exist *Lagrange multipliers* $\boldsymbol{\lambda}^*$ such that a local minimizer $\mathbf{x}^*$ of (12) satisfies:

$$\begin{aligned} \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x}^*,\boldsymbol{\lambda}^*) &= \mathbf{0} \\ c_i(\mathbf{x}^*) &= 0,\ i \in \mathcal{E} \\ c_i(\mathbf{x}^*) &\geqslant 0,\ i \in \mathcal{I} \\ \lambda_i &\geqslant 0,\ i \in \mathcal{I} \\ \lambda_i\, c_i(\mathbf{x}^*) &= 0,\ \forall i. \end{aligned}$$

These equations are usually referred to as the *Karush-Kuhn-Tucker conditions* (KKT). The point $(\mathbf{x}^*,\boldsymbol{\lambda}^*)$ is called a KKT point. The KKT conditions are a necessary condition for a solution to (12). The regularity conditions are rather technical. For a rigorous discussion see Fletcher [20], page 205.

Sufficient conditions for a strict local minimizer are provided by the following theorem:

Theorem 8 (Second order Sufficient Condition) Assume that f and c_i are C^2 functions. Let $\mathcal{H}^* = \nabla^2_{\mathbf{x}}\mathcal{L}(\mathbf{x}^*,\boldsymbol{\lambda}^*)$ be the Hessian matrix of the Lagrangian evaluated at the KKT point $(\mathbf{x}^*,\boldsymbol{\lambda}^*)$. Define the index set of *binding constraints* $\mathscr{A} = \{i :\ c_i(\mathbf{x}^*) = 0\}$ and *strictly active constraints* $\mathscr{A}_+ = \{i :\ i \in \mathcal{E} \text{ or } \lambda_i^* > 0\}$. Let

$$\begin{aligned} \mathscr{G} = \Big\{\mathbf{x} :\ \mathbf{x} \neq 0, \qquad & \nabla_{\mathbf{x}} c_i(\mathbf{x}^*)^T\mathbf{x} = 0, \qquad i \in \mathscr{A} \\ & \nabla_{\mathbf{x}} c_i(\mathbf{x}^*)^T\mathbf{x} \geqslant 0, \qquad i \in \mathscr{A}/\mathscr{A}_+\Big\}. \end{aligned}$$

Then if:

$$\mathbf{x}^T\mathcal{H}^*\mathbf{x} > 0, \quad \forall \mathbf{x} \in \mathscr{G},$$

$\mathbf{x}^*$ is a strict local minimizer of (12).

All the theory above is concerned with finding local solutions to (12). The problem of finding a global minimum is in general very complicated. The concept of convexity, however, gives strong and simple results about the global nature of the solutions of (12).

Definition 5 (Convex Function) Let $\mathbf{x}_\theta = (1-\theta)\mathbf{x}_0 + \theta\mathbf{x}_1$, where $\theta \in [0,1]$ and $\mathbf{x}_1, \mathbf{x}_0 \in \mathscr{X}$. A function f is said to be convex on the set $\mathscr{X}$ if:

$$\mathbf{x}_\theta \in \mathscr{X} \tag{19}$$

$$\text{and} \qquad f(\mathbf{x}_\theta) \leqslant (1-\theta)f(\mathbf{x}_0) + \theta f(\mathbf{x}_1). \tag{20}$$

For $f \in C^1$ the definition implies that f is convex if:

$$f(\mathbf{x}_1) \geqslant f(\mathbf{x}_0) + (\mathbf{x}_1 - \mathbf{x}_2)^T \nabla_{\mathbf{x}} f(\mathbf{x}_0), \quad \forall \mathbf{x}_1, \mathbf{x}_0 \in \mathscr{X}.$$

For $f \in C^2$ the definition implies that f is convex on an open set $\mathscr{X}$ if:

$$\mathbf{x}^T \left[\nabla^2_{\mathbf{x}} f(\mathbf{x})\right] \mathbf{x} \geqslant 0 \quad \forall \mathbf{x} \in \mathscr{X}/\{\mathbf{0}\}.$$

Thus C^2 convex functions are typified by having non-negative curvature. If the inequalities above are strict then f is said to be *strictly convex*[1]. If a function f is (strictly) convex then $-f$ is said to be (strictly) concave. For convex functions we have the following results.

Definition 6 (Convex Programming Problem) The problem of minimizing a convex function f subject to concave constraints c_i on a given set $\mathscr{X}$ is said to be a *convex programming problem*.

Theorem 9 (Convex Optimization) Every local solution $\mathbf{x}^*$ to a convex programing problem is a global solution and the set of global solutions is convex. If, in addition, the objective function is strictly convex, then any global solution is unique.

Theorem 10 (KKT sufficient conditions) For a (strict) convex programming problem with C^1 objective and constraint functions, the KKT conditions are necessary and sufficient for a (unique) global solution.

Duality

The aim of duality is to provide an alternative formulation of an optimization problem which is more computationally convenient or has some theoretical significance (see [20] page 219). The original problem is referred to as the *primal* problem whereas the reformulated problem is referred to as the *dual* problem. Duality theory is most relevant to convex optimization problems. It is well known that if the primal optimization problem is (strictly) convex then the dual problem is (strictly) concave and has a (unique) solution from which the optimal (unique) primal solution can be deduced. In this project we make extensive use the following duality result (see [20] page 219):

Theorem 11 (Wolfe Dual Transformation) Let $\mathbf{x}^*$ be the solution of the convex programming **Primal Problem**:

$$\min_{\mathbf{x} \in \mathbb{R}^d} \; f(\mathbf{x}) \tag{21}$$

$$\text{subject to:} \quad c_i(\mathbf{x}) = 0, \; i \in \mathcal{E} \tag{22}$$

$$c_i(\mathbf{x}) \geqslant 0, \; i \in \mathcal{I} \tag{23}$$

$$f, c_i \in C^1, \tag{24}$$

[1]Note that the converse is also true with the exception that for a strictly convex $f \in C^2$ $\nabla^2_{\mathbf{x}} f(\mathbf{x})$ could be zero. For instance, x^4 is strictly convex yet its second derivative is zero at $x = 0$.

then under mild regularity assumptions there exist Lagrange multipliers $\boldsymbol{\lambda}^*$ such that $\mathbf{x}^*$ and $\boldsymbol{\lambda}^*$ solve the **Dual Problem** :

$$\max_{\mathbf{x},\boldsymbol{\lambda}} \ \mathcal{L}(\mathbf{x},\boldsymbol{\lambda}) \tag{25}$$

$$\text{subject to:} \quad \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{x},\boldsymbol{\lambda}) \ = \ \mathbf{0}, \tag{26}$$

$$\lambda_i \ \geqslant \ 0, \ i \in \mathcal{I}. \tag{27}$$

Furthermore the minimum of the primal and the maximum of the dual function values are equal:

$$f(\mathbf{x}^*) = \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*).$$

Another useful result concerning duality is the following theorem:

Theorem 12 (Duality Gap) Let

$$\min_{\mathbf{x}\in\mathscr{X}} \ f(\mathbf{x}) \tag{28}$$

$$\mathscr{X} \ \equiv \ \{\mathbf{x} : \ c_i(\mathbf{x}) \geqq 0, \ i = 1, \ldots, n\} \tag{29}$$

be a (not necessarily convex) problem with dual:

$$\max_{\{\mathbf{x},\boldsymbol{\lambda}\}\in\Lambda} \ \mathcal{L}(\mathbf{x},\boldsymbol{\lambda}) \tag{30}$$

$$\Lambda \ \equiv \ \{(\mathbf{x},\boldsymbol{\lambda}) : \ \nabla_{\mathbf{x}}\mathcal{L} = \mathbf{0}, \ \boldsymbol{\lambda} \geqslant \mathbf{0}\} \tag{31}$$

Then

$$v = \inf_{\mathbf{x}\in\mathscr{X}} \ f(\mathbf{x}) \geqslant \omega = \sup_{\{\mathbf{x},\boldsymbol{\lambda}\}\in\Lambda} \ \mathcal{L}(\mathbf{x},\boldsymbol{\lambda}).$$

The difference $v - \omega$ is called the *duality gap*.

The Duality Gap theorem is extremely useful for providing lower bounds to the solutions of primal problems which may by impossible to solve directly. For convex programming problems the duality gap is zero. This property is usually referred to as *strong duality*. Sometimes, however, the primal problem may be unbounded ($f \to -\infty$) in which case by the Duality Gap theorem $v = \omega = -\infty$. Hence an unbounded primal implies an inconsistent dual. In this project we will be dealing primarily with linearly constrained programming problems so it important to note that for linearly constrained problems, if the primal is infeasible (does not have a solution satisfying the constraints), then the dual is either infeasible or unbounded. Conversely if the dual is infeasible then the primal has no solution.

Example 6 (LPP) Consider the LPP in standard form:

$$\min_{\mathbf{x}} \ c_0 + \mathbf{c}^T\mathbf{x} \tag{32}$$

$$\text{subject to:} \quad A\mathbf{x} \ \geqslant \ \mathbf{b}. \tag{33}$$

Since the objective function is convex and the constraints are concave, application of the Wolfe dual gives:

$$\max_{\lambda} \; c_0 + \mathbf{b}^T\boldsymbol{\lambda} \tag{34}$$

$$\text{subject to:} \quad A\boldsymbol{\lambda} = \mathbf{c}, \tag{35}$$

$$\boldsymbol{\lambda} \geqslant \mathbf{0}. \tag{36}$$

It is interesting to note that for the LPP the dual of the dual problem always gives back the primal problem.

Of more interest in the project is the application of the duality transformation to the QPP.

Example 7 (QPP) Consider the QPP:

$$\min_{\mathbf{x}} \; \frac{1}{2}\mathbf{x}^T C\mathbf{x} - \frac{1}{2} \tag{37}$$

$$\text{subject to:} \quad C\mathbf{x} \geqslant \boldsymbol{\kappa}^*, \tag{38}$$

where the matrix $C_{n\times d}$ is positive definite. Again, since the objective function is convex and the constraints are concave, the problem is a convex programming problem. We can thus write the dual problem:

$$\max_{\mathbf{x}} \quad -\frac{1}{2}\mathbf{x}^T C\mathbf{x} + \mathbf{x}^T\boldsymbol{\kappa}^* - \frac{1}{2} \tag{39}$$

$$\text{subject to:} \quad \mathbf{x} \geqslant \mathbf{0}. \tag{40}$$

Notice that the dual problem involves only simple inequality *box constraints*. This could possibly make it easier to solve than the primal problem[2]. We thus have a choice as to which one of the problem formulations we choose to solve numerically. Sometimes this choice is important because the two problems differ in their numerical properties. This is especially important if C is numerically ill-conditioned. For example, a conjugate gradient trust region algorithm (see [15]) applied to the box constrained formulation may take many iterations to converge. For ill conditioned matrices an alternative possibility is to maintain primal-dual feasibility by optimizing not just the primal or the dual formulation but both of them simultaneously. The idea is to minimize the duality gap between the primal and the dual objective function:

$$\min_{\lambda} \quad \boldsymbol{\lambda}^T C\boldsymbol{\lambda} - \boldsymbol{\lambda}^T\boldsymbol{\kappa}^* = \boldsymbol{\lambda}^T(C\boldsymbol{\lambda} - \boldsymbol{\kappa}^*) = \text{Duality Gap} \tag{41}$$

$$\text{subject to:} \quad C\boldsymbol{\lambda} \geqslant \boldsymbol{\kappa}^*, \;\; \boldsymbol{\lambda} \geqslant \mathbf{0} \tag{42}$$

[2]The optimization literature seems to be saturated with various large scale algorithms for the solution of the box constrained QPP problem.

Since the problem is strictly convex we know that at the optimal solution the duality gap must be zero. This implies the complementarity condition, i.e., either $\sum_k C_{ik}\lambda_k = k_i^*$ or $\lambda_i = 0$ at the optimal solution. Minimization of the duality gap involves more computation since there are more constraints but it has been observed to be stable for large ill-conditioned matrices.

Example 8 (QPP with Cholesky factorization) Now suppose that we are given the Cholesky factorization $C = L^T L$. Then setting $\boldsymbol{\mu} = L\boldsymbol{\lambda}$, the primal becomes:

$$\min_{\boldsymbol{\mu}} \quad \frac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\mu} \tag{43}$$

$$\text{subject to:} \quad L^T\boldsymbol{\mu} \geqslant \boldsymbol{\kappa}^*. \tag{44}$$

This is a so called *least distance* problem which, provided we know the Cholesky factorization of C, is easier to solve than the original QPP.

A final example of duality is provided by the widely used Maximum Entropy method [26].

Example 9 (GPP) Suppose we are given the primal GPP:

$$\min_{\mathbf{p}} \quad \sum_{m=1}^{M} p_m \ln \frac{p_m}{q_m} \tag{45}$$

$$\text{subject to:} \quad \mathbf{p} \geqslant \mathbf{0},\ \sum_{m=1}^{M} p_m K_{i,m} \geqslant \kappa_i^*,\ \ i = 1, \ldots, n \tag{46}$$

where $n \lll M$ and $q_m > 0$. Here the objective function is a linear combination of functions of the form $x\ln(x/c)$. These functions are convex for $x \in \mathbb{R}^+$ and $c > 0$. A linear combination of convex functions is another convex function. Hence we have a convex programming problem. The Lagrangian is $\mathcal{L}(\mathbf{p}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \sum_{m=1}^{M} p_m \ln \frac{p_m}{q_m} - \sum_{i=1}^{n} \lambda_i \left(\sum_{m=1}^{M} p_m K_{i,m} - \kappa_i^*\right) - \sum_{m=1}^{M} \mu_m p_m$ and the dual is :

$$\max_{\mathbf{p},\boldsymbol{\lambda},\boldsymbol{\mu}} \quad \mathcal{L}(\mathbf{p}, \boldsymbol{\lambda}, \boldsymbol{\mu}) \tag{47}$$

$$\text{subject to:} \quad \ln \frac{p_m}{q_m} = -1 + \mu_m + \sum_{i=1}^{n} \lambda_i K_{i,m} \tag{48}$$

$$\boldsymbol{\lambda} \geqslant \mathbf{0},\ \boldsymbol{\mu} \geqslant \mathbf{0} \tag{49}$$

Therefore $p_m = q_m \exp\left(-1 + \mu_m + \sum_{i=1}^{n} \lambda_i K_{i,m}\right)$, which is always non-negative. Thus the constraint $\mathbf{p} \geqslant \mathbf{0}$ is inactive and $\boldsymbol{\mu} = \mathbf{0}$. Eliminating $\mathbf{p}$ from the Lagrangian gives the dual:

$$\max_{\boldsymbol{\lambda}} \quad \boldsymbol{\lambda}^T\boldsymbol{\kappa}^* - \sum_{m=1}^{M} q_m \exp\left(\sum_{i=1}^{n} \lambda_i K_{i,m} - 1\right) \tag{50}$$

$$\text{subject to:} \quad \boldsymbol{\lambda} \geqslant \mathbf{0} \tag{51}$$

Note that the dual problem involves only n variables and is thus easier to solve than the primal. In fact M can be so much larger than n that the only possible way of obtaining a solution to the primal is via the dual problem. We will exploit this property when applying the GCE method to discrete spaces of large cardinality.

The duality theory discussed above for a finite dimensional convex programming problem of the form (12) extends to infinite dimensional functional optimization problems in which the integrand is convex and the constraints are linear. For more details on this quite technical issue see [67] page 219, Decarreau [1], Borwein [7] and the references therein. A simple application of the duality theory for functional optimization problems is given in the description of the GCE method for continuous optimization.

1.6 Statistical Learning Theory

As was stated earlier, the generic problem of Statistical Learning Theory is to find/estimate an optimal (in some sense) probability function from a finite number of empirical observations. We now briefly review some results concerning this problem.

Estimating the Distribution Function

Suppose we are given data $\mathcal{X}_n = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}, \quad \mathbf{x}_i \in \mathbb{R}^d$ and wish to estimate its distribution function $F : \mathbb{R}^d \to [0, 1]$. Then an estimate of the unknown distribution function can be

$$\hat{F}_n(\mathbf{x}) = \hat{F}(\mathbf{x} \,|\, \mathcal{X}_n) = \frac{\#\{\mathbf{X}_i \leqslant \mathbf{x}\}}{n} = \frac{1}{n} \sum_{i=1}^{n} I\{\mathbf{X}_i \leqslant \mathbf{x}\}.$$

The inequality $\{\mathbf{X}_i \leqslant \mathbf{x}\}$ is applied component-wise. The $\hat{F}_n$ denotes the fact that the estimate depends on the number of observations. Scott [56] gives the following result concerning the estimator of F:

Theorem 13 (MVUE of F) The estimator $\hat{F}_n(\mathbf{x})$ is the minimum variance unbiased estimator (MVUE) of $F(\mathbf{x})$.

The result follows from the fact that $\hat{F}_n$ is both unbiased and a function of the order statistics which form a complete sufficient statistic (see Pawitan [49]). Notice however that $\hat{F}_n$ is always piecewise continuous even when F is known to be smooth and continuously differentiable.

Estimating The Density Function

Many density estimators based on the empirical distribution can be written as a linear combination of localized functions. This includes estimators based on orthogonal series expansions, splines and estimators which are solutions to functional regularization problems. This observation is known as the General Kernel Theorem [56], page 156. To introduce the theorem we need:

Definition 7 (Gateaux Derivative) The Gateaux derivative of a functional J at the function ϕ in the direction of the function η is defined to be :

$$\mathcal{G}\{\phi\}(\eta) = \lim_{\|\varepsilon\| \to 0^+} \frac{J[\phi + \varepsilon\eta] - J[\phi]}{\|\varepsilon\|}.$$

Theorem 14 (General Kernel Theorem) Any density estimator that is a continuous and Gateaux differentiable functional of the empirical distribution may be written as

$$f_n(\mathbf{x}) = f(\mathbf{x} \,|\, \mathcal{X}_n) = \frac{1}{n} \sum_{i=1}^{n} K(\mathbf{x}, \mathbf{X}_i, \hat{F}_n), \tag{52}$$

where K is the Gateaux derivative of f_n under variation of $\mathbf{X}_i$. Thus K , which is called a kernel function, measures the influence of $\mathbf{X}_i$ on f_n.

Proof: Consider the one-dimensional case. We can then write the distribution function and the density estimator as an operator:

$$\begin{aligned} \hat{F}_n(\cdot) &= \frac{1}{n} \sum_{i=1}^{n} I_{[X_i,\infty)}(\cdot) \\ \text{and} \qquad f(x \,|\, \mathcal{X}_n) &= T_x\left\{\hat{F}_n\right\}. \end{aligned}$$

Then define:

$$\begin{aligned} K(x; X_i, \hat{F}_n) &= \lim_{\varepsilon \to 0} \frac{T_x\left\{(1-\varepsilon)\hat{F}_n + \varepsilon\, I_{[X_i,\infty)}\right\} - T_x\left\{(1-\varepsilon)\hat{F}_n\right\}}{\varepsilon} \\ &= T_x\left\{\hat{F}_n\right\} + \lim_{\varepsilon \to 0} \frac{T_x\left\{\hat{F}_n + \varepsilon\left(I_{[X_i,\infty)} - \hat{F}_n\right)\right\} - T_x\left\{\hat{F}_n\right\}}{\varepsilon} \\ &= T_x\left\{\hat{F}_n\right\} + \mathcal{G}\left\{\hat{F}_n\right\}\left(I_{[X_i,\infty)} - \hat{F}_n\right). \end{aligned}$$

By linearity of the Gateaux derivative we have:

$$\begin{aligned} \frac{1}{n}\sum_{i=1} K(x; X_i, \hat{F}_n) &= T_x\left\{\hat{F}_n\right\} + \frac{1}{n}\sum_{i=1}^{n} \mathcal{G}\left\{\hat{F}_n\right\}\left(I_{[X_i,\infty)} - \hat{F}_n\right) \\ &= T_x\left\{\hat{F}_n\right\} + \mathcal{G}\left\{\hat{F}_n\right\}\left(\frac{1}{n}\sum_{i=1}^{n} I_{[X_i,\infty)} - \hat{F}_n\right) \\ &= T_x\left\{\hat{F}_n\right\} + \mathcal{G}\left\{\hat{F}_n\right\}(0) = T_x\left\{\hat{F}_n\right\} = f(x \,|\, \mathcal{X}_n). \end{aligned}$$

This concludes the proof.

Let q^* be the density function associated with F. From the definition of the density function as the derivative of the distribution function, we obtain the unbiased estimator of q^*:

$$f_n(\mathbf{x}) = f(\mathbf{x} \mid \mathcal{X}_n) = \nabla_{\mathbf{x}} \hat{F}_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \delta(\mathbf{x} - \mathbf{X}_i),$$

where $\delta(\mathbf{x})$ is the multidimensional Dirac Delta function. Although this estimator fits the General Kernel Theorem and is unbiased, it has infinite variance[3] and is useless when the underlying true q^* is known to be a PWC function. Unfortunately, while a MVUE of the distribution function F exists, for the density we have the following result proved by Rosenblatt [52]:

Theorem 15 (MVUE of Density) Suppose $\mathcal{X}_n = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ is a random sample from the continuous density $q^*(\mathbf{x})$. Then for **any** estimator $f(\mathbf{x} \mid \mathcal{X}_n)$

$$\mathbb{E}_{q^*}[f(\mathbf{x} \mid \mathcal{X}_n)] = q^*(\mathbf{x}), \quad \forall n, \mathbf{x}$$

is impossible. In other words there does not exist a finite variance unbiased estimator of the density function $q^*(\mathbf{x})$.

Proof: Consider the one-dimensional case. Assume that $\mathbb{E}_{q^*}[f(x \mid \mathcal{X}_n)] = q^*(x), \ \forall n, x$, is possible, then by Fubini's theorem

$$\begin{aligned} \mathbb{E}_{q^*}\left[\int_a^b f(x \mid \mathcal{X}_n)\, dx\right] &= \int_a^b \mathbb{E}_{q^*}\left[f(x \mid \mathcal{X}_n)\right] \\ &= \int_a^b q^*(x)\, dx \\ &= F(b) - F(a) \\ &= \mathbb{E}_{q^*}\left[\hat{F}_n(b) - \hat{F}_n(a)\right] \end{aligned}$$

Both $f(x \mid \mathcal{X}_n)$ and $\hat{F}_n(b) - \hat{F}_n(a)$ are functions of the complete sufficient statistic and since $\hat{F}_n(b) - \hat{F}_n(a)$ is the only symmetric unbiased estimator of $F(b) - F(a)$ we must have

$$\hat{F}_n(b) - \hat{F}_n(a) = \int_a^b f(x \mid \mathcal{X}_n)\, dx, \quad \forall \mathcal{X}_n.$$

This is impossible since the right-hand side is absolutely continuous whereas the left-hand side is not.

One way out of this predicament is to require an estimator with finite variance and asymptotic unbiasedness. This requirement gives rise to the *non-parametric* estimators discussed next.

[3] Consider $\mathbb{E}_q[\delta^2(\mathbf{X} - \mathbf{X}_i)] = \int q(\mathbf{x})\, \delta(\mathbf{x} - \mathbf{X}_i)\, \delta(\mathbf{x} - \mathbf{X}_i)\, d\mathbf{x} = q(\mathbf{X}_i)\, \delta(\mathbf{X}_i - \mathbf{X}_i)$, which is an infinite spike for $q(\mathbf{X}_i) > 0$.

Non-parametric density estimators

A parametric estimator of q^* is defined by any parametric model $f(\mathbf{x}, \boldsymbol{\theta} \mid \mathcal{X}_n)$ with parameter $\boldsymbol{\theta} \in \Theta$, where the dimension of Θ is fixed and constant for any sample size. An intuitive definition of non-parametric estimators is an estimator with infinite number of parameters or a number of parameters which diverges as the sample size diverges. Alternatively, for nonparametric estimators, if $\|\mathbf{x}-\mathbf{X}_i\| > \varepsilon$ for any $\varepsilon > 0$, the influence of the data point $\mathbf{X}_i$ on the point density estimate at $\mathbf{x}$ vanishes asymptotically. In other words the influence of the sample points outside an $\varepsilon-$neighborhood of $\mathbf{x}$ must vanish as $n \to \infty$. Thus non-parametric estimators are asymptotically local, while parametric estimators are not. Note, however, that $\frac{1}{n}\sum_{i=1}^{n} \delta(\mathbf{x} - \mathbf{X}_i)$ is asymptotically local yet useless as an estimator of a PWC density. This problem is avoided by insisting that non-parametric estimators be consistent.

Definition 8 (Consistency of Density Estimators) A density estimator f_n of q^* is said to be consistent if:

$$\lim_{n\to\infty} \mathrm{MSE}\left\{f_n\right\}(\mathbf{x}) = 0, \quad \forall \mathbf{x} \in \mathbb{R}^d,$$

where $\mathrm{MSE}\left\{f_n\right\}(\mathbf{x}) = \mathbb{E}_{q^*}\left[f(\mathbf{x} \mid \mathcal{X}_n) - q^*(\mathbf{x})\right]^2 = \mathrm{Var}_{q^*}\left[f_n(\mathbf{x})\right] + \mathrm{Bias}^2_{q^*}\left[f_n(\mathbf{x})\right]$ is the *Mean Squared Error* of f_n at the point $\mathbf{x}$.

Definition 9 (Non-parametric Density Estimator) A density estimator f_n is said to be *non-parametric* when f_n is consistent in the Mean Squared Error sense.

The condition of consistency of f_n can be translated into specific requirements on the kernel functions $K(\mathbf{x}, \mathbf{X}_i, \hat{F}_n)$ given below (see [56], page 157).

Theorem 16 (General Kernel Density Estimator) Let $f_n(\mathbf{x})$ be a continuous and Gateaux differentiable density estimator based on the empirical distribution function, i.e., it can be written as $f_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} K(\mathbf{x}, \mathbf{X}_i, F_n)$. Then $f_n(\mathbf{x})$ is a non-parametric density estimator provided:

1.
$$\lim_{n\to\infty} \int K(\mathbf{x}, \mathbf{X}_i, \hat{F}_n)\, d\mathbf{x} = 1,$$

2.
$$\lim_{n\to\infty} \int \mathbf{x}\, K(\mathbf{x}, \mathbf{X}_i, \hat{F}_n)\, d\mathbf{x} = \mathbf{X}_i,$$

i.e., $\lim_{n\to\infty} K(\mathbf{x}, \mathbf{X}_i, \hat{F}_n) = \delta(\mathbf{x} - \mathbf{X}_i)$;

3.

$$\lim_{n\to\infty} \Sigma_{\mathbf{X}_i,n} = \mathbf{0},$$

$$\lim_{n\to\infty} n\ \Sigma_{\mathbf{X}_i,n} = \infty,$$

where $\Sigma_{\mathbf{X}_i,n} = \int (\mathbf{x} - \mathbf{X}_i)(\mathbf{x} - \mathbf{X}_i)^T K(\mathbf{x}, \mathbf{X}_i, \hat{F}_n)\ d\mathbf{x} \neq \mathbf{0}, \quad \forall n.$

We will come back to the problems of non-parametric statistics in the last section.

2 The Cross Entropy Postulate

We now describe a generic version of the GCE method. The GCE method is related to the CE method [54] and the Generalized Entropy Optimization Principles presented in [33]. Similar to the CE method the GCE associates a *proposal* probability density with the problems of Monte Carlo simulation and Machine Learning. This density is then iteratively updated in view of the observed empirical behavior of the resulting probabilistic system. The updating aims to "steer" the instrumental density toward an optimal (in some sense) density — the *target* density. Knowledge of the target density usually equates to knowledge of the solution of the original problem. For this purpose we need a mechanism for updating a given probability density in view of incoming information about the observed probabilistic system. One such consistent and axiomatically rigorous mechanism is provided by the *Cross Entropy Postulate* (see [33] and [30]).

Definition 10 (The Cross Entropy Postulate) *Given any three of the probabilistic entities:*

1. *an a-priori probability density q,*
2. *a generalized Cross Entropy distance $\mathcal{D}$ (also known as relative/directed divergence) between two probability densities,*
3. *a set $\mathscr{C}$ of constraints connecting the probabilistic entities with the observed behavior of the system,*
4. *an a-posteriori density p,*

then under suitable conditions the fourth entity can be found uniquely.

The postulate is important for the correct interpretation of the GCE method. It provides a consistent and self-sufficient framework for inference and a mechanism for updating a given probability model in view of newly available information. We need to specify three of the probabilistic entities to use the postulate.

2.1 The Prior Probability Density

The GCE method assumes that the proposal probability density is updated iteratively. The a-priori density q at the current iteration is the a-posteriori density from the previous iteration. The a-priori density which is used to initialize the iteration is the uniform density over the region of interest. In some cases the prior density is the improper uniform density and the normalizing constant over the region of interest is not strictly computable. Similar to the Bayesian methodology the GCE takes $q(\mathbf{x}) \propto 1,\ \forall \mathbf{x} \in \mathscr{X}$ without any reference to the value of the normalizing constant. The GCE always takes the uniform density, *improper* or otherwise, as the most unbiased and uninformative prior density[4]. This is in accordance with Laplace's *Principle of Insufficient Reason* [38], which argues that the uniform density is the most unbiased and objective density in the absence of any information about the analyzed probabilistic system. In cases where we use the improper prior $q \propto 1$ the method is similar in nature to the Maximum Entropy Method (MEM) of Jaynes [26].

2.2 The Cross Entropy distance $\mathcal{D}$

We use the notion of Cross Entropy distance (directed divergence) between two probability densities. We restrict our attention to the class of directed divergence measures first analyzed by Csiszár [16]. These measures constitute a direct generalization of the most widely used and computationally tractable information-theoretic measures since the birth of Information Theory [58]. A distinguishing property of these measures is their convexity.

Definition 11 (Csiszár Measure) *The Csiszár generalized measure of directed divergence between two continuous probability densities p and q is:*

$$\mathcal{D}(p \to q) \quad = \quad \int q(\mathbf{x})\ \psi\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right) d\mathbf{x}\ , \quad \mathbf{x} \in \mathbb{R}^d,$$

where

1. $\psi :\ \mathbb{R}^+ \to \mathbb{R}$ *is a continuous twice-differentiable function;*
2. $\psi(1) = 0$*;*
3. $\psi''(x) > 0$ *for all* $x \in \mathbb{R}^+$.

[4]This is in contrast to the Bayesian approach which uses the so called uninformative *Jeffrey's priors* — densities defined over the space Θ of a model parameter $\boldsymbol{\theta}$ and usually very different from the uniform density over the set $\mathscr{X}$. In the GCE method we deal directly with the most uninformative density over the space of the observables, i.e., the uniform density over $\mathscr{X}$.

There are no conceptual differences for the case in which p and q are discrete densities. The integral is simply replaced by the sum: $\sum_i q_i \, \psi\left(\frac{p_i}{q_i}\right)$.

The definition of the Csiszár's measure ensures that $\mathcal{D}$ has the properties:

1. $\mathcal{D}(p \to q) \geqslant 0$ following Jensen's inequality $\mathbb{E}_q \psi\left(\frac{p(\mathbf{X})}{q(\mathbf{X})}\right) \geqslant \psi\left(\mathbb{E}_q \frac{p(\mathbf{X})}{q(\mathbf{X})}\right) = \psi(1)$.
2. $\mathcal{D}(p \to q) = 0$ if and only if $p \equiv q$.
3. In the discrete case $\mathcal{D}$ is permutationally symmetric, i.e., it does not change when the pairs $(p_1, q_1), (p_2, q_2), \ldots, (p_n, q_n)$ are permuted amongst themselves.
4. $\mathcal{D}(p \to q)$ is a convex function of p and q.
5. $\mathcal{D}$ is continuous and differentiable with respect to p and q.

Properties 1, 2 and 3 are essential for any meaningful measure of distance. Properties 4 and 5 are important in ensuring mathematical tractability when using the measure in practical optimization problems. We can think of $\mathcal{D}$ as measuring the divergence/distance of p from q in some appropriate probability space. Notice however that $\mathcal{D}$ is not a distance in the usual Euclidian sense:

- in general $\mathcal{D}(p \to q) \neq \mathcal{D}(q \to p)$, i.e., $\mathcal{D}$ is not symmetric;
- in general $\mathcal{D}(p \to q) + \mathcal{D}(q \to s) \ngeq \mathcal{D}(p \to s)$ for any probability density s, i.e., the measure does not satisfy the triangle inequality which is characteristic for all Euclidian measures of distance.

Csiszár's family of measures subsumes all of the information-theoretic measures used in practice (see [6], [32], [59] and [4]). To see this set

$$\psi(x) = \frac{x^\alpha - x}{\alpha(\alpha - 1)}, \quad \alpha \neq 0, 1.$$

The polynomial $\frac{x^\alpha - x}{\alpha(\alpha-1)}$ is the simplest differentiable function satisfying the conditions $\psi(1) = 0$ and $\psi''(x) > 0$ for $x > 0$. The resulting CE distance:

$$\mathcal{D}_\alpha(p \to q) = \frac{1}{\alpha(\alpha - 1)} \left(\int p^\alpha(\mathbf{x}) q^{1-\alpha}(\mathbf{x}) \, d\mathbf{x} - 1 \right) \tag{53}$$

is indexed by the parameter α. This parametric family of CE measures was first studied by Havrda-Charvat [25]. Specific choices of α give rise to the most notable CE measures:

1. *Hellinger distance*:

$$\mathcal{D}_{1/2}(p \to q) = 2 \int \left(\sqrt{p(\mathbf{x})} - \sqrt{q(\mathbf{x})} \right)^2 d\mathbf{x} \tag{54}$$

Note the symmetry $\mathcal{D}_{1/2}(p \to q) = \mathcal{D}_{1/2}(q \to p)$ of this particular member of the Csiszár family.

2. *Pearson χ^2 discrepancy measure*:

$$\mathcal{D}_2(p \to q) = \frac{1}{2} \int \frac{\left[p(\mathbf{x}) - q(\mathbf{x})\right]^2}{q(\mathbf{x})} \, d\mathbf{x} \tag{55}$$

3. *Neymann χ^2 'goodness of fit' measure*:

$$\mathcal{D}_{-1}(p \to q) = \frac{1}{2} \int \frac{\left[p(\mathbf{x}) - q(\mathbf{x})\right]^2}{p(\mathbf{x})} \, d\mathbf{x} \tag{56}$$

4. *Burg CE distance* [11]:

$$\lim_{\alpha \to 0} \mathcal{D}_\alpha(p \to q) = \int q(\mathbf{x}) \ln \left(\frac{q(\mathbf{x})}{p(\mathbf{x})} \right) d\mathbf{x} \tag{57}$$

5. *Kullback-Leibler CE distance* [42]:

$$\lim_{\alpha \to 1} \mathcal{D}_\alpha(p \to q) = \int p(\mathbf{x}) \ln \left(\frac{p(\mathbf{x})}{q(\mathbf{x})} \right) d\mathbf{x} \tag{58}$$

Remark 1 The relation

$$\mathcal{D}_\alpha(p \to q) = \mathcal{D}_{1-\alpha}(q \to p)$$

holds for all α including the special limiting cases for $\alpha \to 1$ or $\alpha \to 0$.

Remark 2 For optimization purposes it does not matter whether we use $\mathcal{D}(p \to q)$ or $F(\mathcal{D}(p \to q))$ where F is a monotonic function; For example, minimizing the *Renyi* CE distance [51]:

$$\min_p \quad \frac{1}{\alpha - 1} \ln \left(\int p^\alpha(\mathbf{x}) q^{1-\alpha}(\mathbf{x}) \, d\mathbf{x} \right), \qquad \alpha > 0 \quad \alpha \neq 1$$

gives the same result as

$$\min_p \quad \frac{1}{\alpha(\alpha - 1)} \int p^\alpha(\mathbf{x}) q^{1-\alpha}(\mathbf{x}) \, d\mathbf{x}, \qquad \alpha > 0 \quad \alpha \neq 1.$$

In fact [32] argues that *Renyi* and *Havrda-Charvat* CE measures are equivalent in the sense that when they are maximized (minimized) the resulting maximizing (minimizing) probability densities are the same.

A useful relation between Pearson's χ^2 measure and Kullback-Leibler CE measure is (see Devroye [17], page 224):

$$\int \frac{[p(\mathbf{x}) - q(\mathbf{x})]^2}{q(\mathbf{x})}\, d\mathbf{x} \geqslant \int p(\mathbf{x})\ \ln \frac{p(\mathbf{x})}{q(\mathbf{x})} \geqslant \ln\left(1 + \int \frac{[p(\mathbf{x}) - q(\mathbf{x})]^2}{q(\mathbf{x})}\, d\mathbf{x}\right),$$

i.e.,

$$2\,\mathcal{D}_2(p \to q) \geqslant \lim_{\alpha \to 1} \mathcal{D}_\alpha(p \to q) \geqslant \ln\left(1 + 2\mathcal{D}_2(p \to q)\right).$$

It is also easy to show that $\mathcal{D}_2$ is related to the L_1 distance.

Lemma 2 (Relation to L_1 metric)

$$2\mathcal{D}_2(p \to q) \geqslant \left(\int |p(\mathbf{x}) - q(\mathbf{x})|\, d\mathbf{x}\right)^2.$$

Proof: From the properties of the Csiszár measure we know that $\int \frac{p^2}{q}\, d\mathbf{x} \geqslant 1$ *for any two non-negative functions p and q which integrate to one. More specifically we can have arbitrary non-negative functions f and g , which do not necessarily integrate to one, and for which we have:* $\int \frac{f^2}{g}\, d\mathbf{x} \geqslant \frac{\left(\int f\, d\mathbf{x}\right)^2}{\int g\, d\mathbf{x}}$. *Now let* $f(\mathbf{x}) = |a(\mathbf{x}) - b(\mathbf{x})|$ *for two arbitrary density functions a and b. Then setting* $g = a$*, we obtain*

$$\int \frac{(a-b)^2}{a}\, d\mathbf{x} \geqslant \left(\int |a - b|\, d\mathbf{x}\right)^2.$$

In later sections through a long and twisted argument we will arrive at the $\mathcal{D}_2$ measure and show that the advantage of $\mathcal{D}_2$ over all the other measures is its computational tractability and ease of interpretation. Another advantage of $\mathcal{D}_2$ is that by minimizing the $\mathcal{D}_2$ distance we minimize an upper bound on two very fundamental metrics —the Kullback-Leibler [41] and L_1 measures. For example the L_1 metric is the only L_p metric that is invariant to monotone transformations of $\mathbf{x}$. Devroye [17] has written a whole book on the theoretical significance of the L_1 metric in the context of density estimation.

Another important property of $\mathcal{D}_2$ is that it is an approximation to the Kullback-Leibler CE measure. To see this consider the discrete case with $\mathbf{q} = (q_1, \ldots, q_M)$ and $\mathbf{p} = (p_1 \ldots, p_M)$. Suppose we can write $p_i = q_i(1 + \varepsilon_i),\ \forall i$ for

some not very large perturbation $|\varepsilon_i| < 1$ with $\mathbb{E}_{\mathbf{q}}[\varepsilon] = 0$, then:

$$\begin{aligned}
\sum_i p_i \ln \frac{p_i}{q_i} &= \sum_i q_i(1+\varepsilon_i)\ln(1+\varepsilon_i) \\
&= \sum_i q_i(1+\varepsilon_i)\left(\varepsilon_i - \frac{\varepsilon_i^2}{2} + \frac{\varepsilon_i^3}{3} - \cdots\right), \quad \text{for } |\varepsilon_i| < 1 \\
&= \sum_i q_i\left(\varepsilon_i + \frac{\varepsilon_i^2}{2} + \text{higher order terms}\right) \\
&\approx \sum_i q_i\left(\varepsilon_i + \frac{\varepsilon_i^2}{2}\right) = \frac{1}{2}\sum_i \frac{(p_i - q_i)^2}{q_i} \\
&= \mathcal{D}_2(\mathbf{p} \to \mathbf{q}) = \frac{1}{2}\mathrm{Var}_{\mathbf{q}}(\varepsilon).
\end{aligned}$$

Now that we have a reasonable choice for the second ingredient of the CE postulate we comment briefly on the third ingredient.

2.3 The Constraint Set $\mathscr{C}$

For the purposes of the GCE method the density p is required to satisfy a finite number of integral constraints of the from:

$$\mathbb{E}_p K_i(\mathbf{X}) \gtreqless \mathbb{E}_{q^*} K_i(\mathbf{X}), \quad i = 1, \ldots, n,$$

where $\{K_i\}_{i=1}^n$ is a set of suitably chosen functions and q^* is the density which solves a statistical learning or simulation problem. For example each K_i can be a Gaussian density and q^* can be the optimal Importance Sampling density for rare-event simulation (see [54]). Note that the CE postulate gives us a consistent updating rule when any three of the probabilistic entities have been chosen. It does not, however, provide any guidance as to the choice of the probabilistic entities in the first place. Our choice of $\mathscr{C}$ will be guided by the results of Statistical Learning theory and the following considerations:

1. If the expectations $\mathbb{E}_{q^*} K_i(\mathbf{X})$ have to be estimated from empirical data then the corresponding estimators κ_i^* should be (asymptotically) efficient. I.e., κ_i^* should preferably be the Maximum Likelihood Estimator of $\mathbb{E}_{q^*} K_i(\mathbf{X})$.

2. The computation of κ_i^* should be easy. For example a computationally manageable and reliable estimate of $\mathbb{E}_{q^*} K_i(\mathbf{X})$ may be the Monte Carlo average $\kappa_i^* = \frac{1}{J}\sum_{j=1}^J K_i(\mathbf{X}_j)$, where $\mathbf{X}_1, \ldots, \mathbf{X}_J \sim q^*$.

The constraints in $\mathscr{C}$ are linear integral constraints. Concerning the constraint set $\mathscr{C}$ we have the following definition (see [38]):

Definition 12 (Characterizing moments) Suppose that the CE distance $\mathcal{D}$, the a-priori density q and the constraint set $\mathscr{C}$ are specified in the CE postulate. Suppose further that the posterior density p can be derived from the CE postulate and is unique. Then p is said to be *characterized* by the constraint set $\mathscr{C}$ under the CE measure $\mathcal{D}$ and the a-priori density q. The constraints in $\mathscr{C}$ are referred to as *characterizing* constraints of the density p. Moreover, if the constraints are linear and integral, then they are said to be the *charactering moment* constraints of the density p.

The constraints connecting the probabilistic model with the observed behavior of the system embody nothing more than a generalization of the *moment matching* idea of Karl Pearson. We match the characterizing moments of the proposed model $\mathbb{E}_p[K_i(\mathbf{X})]$ to the empirical moments κ_i^* (which approximate the true but unknown $\mathbb{E}_{q^*}[K_i(\mathbf{X})]$). We are now ready to combine the three specified ingredients and apply the postulate to obtain the fourth probabilistic entity, i.e., the posterior density p.

3 A generic GCE algorithm

In this section a quite general iterative algorithm for stochastic optimization and machine learning is presented. Suppose that at a given step of the iterative procedure we have a given a-priori proposal sampling density q which we wish to update on the basis of empirical data with the aim of learning more about the unknown stochastic process. Furthermore let the target density which solves the simulation, optimization or learning problem be denoted as q^* (e.g., q^* could be the optimal Importance Sampling density). Then the a-priori density q is updated to p using the CE postulate with the following ingredients:

1. Given the a-priori probability density q on the set $\mathscr{X} \subset \mathbb{R}^d$,
2. minimize the Csiszár measure of Cross Entropy :

$$\mathcal{D}(p \to q) = \int_{\mathscr{X}} q(\mathbf{x})\ \psi\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right) d\mathbf{x} \tag{59}$$

in terms of the density p, where $\mathbf{x} \in \mathbb{R}^d$ is a column vector. In other words we have to solve the functional optimization problem:

$$\min_{p \in \mathscr{P}} \mathcal{D}(p \to q), \tag{60}$$

where $\mathscr{P} = \left\{p :\ \int p(\mathbf{x})\ d\mathbf{x} = 1,\ p(\mathbf{x}) \geqslant 0, \forall \mathbf{x} \in \mathscr{X}\right\}$ is the set of all bona fide density functions on $\mathscr{X}$,

3. subject to the *characterizing moment* constraints:

$$\mathbb{E}_p K_i(\mathbf{X}) = \int_{\mathscr{X}} p(\mathbf{x})\, K_i(\mathbf{x})\, d\mathbf{x} \;\; \geqslant \;\; \kappa_i^* \,, \quad i = 1, \ldots, n, \tag{61}$$

where

a) κ_i^* is a stochastic or deterministic estimate of $\mathbb{E}_{q^*} K_i(\mathbf{X})$,

b) each $K_i : \mathbb{R}^d \to \mathbb{R}$ is an absolutely continuous function. The K_i's are usually referred to as *kernel* functions. Typically the GCE method assumes that each kernel K_i has the properties:

1. $\int_{\mathscr{X}} K_i(\mathbf{x})\, d\mathbf{x} = 1, \;\; K_i(\mathbf{x}) \geqslant 0$ for all $\mathbf{x} \in \mathbb{R}^d$,

2. $K_i(\mathbf{x}) = K_i(-\mathbf{x})$, i.e., the kernel is an even/symmetric function,

3. $K_i(\mathbf{x}) = K(\mathbf{x}; \mathbf{x}_i, \Sigma_i)$, so that each kernel K_i has a fixed functional form but variable location and shape parameters $\mathbf{x}_i$ and Σ_i respectively. The location parameters $\mathcal{X}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ are (usually independent and identically distributed) column vector realizations from the a-priori density q or if possible from the target q^*. Each Σ_i is a symmetric $d \times d$ positive definite matrix. Σ is usually referred to as the *bandwidth* or *scale* matrix of the kernel K. For example,

$$K_i(\mathbf{x}) = K(\mathbf{x}; \mathbf{x}_i, \Sigma_i) = |\Sigma_i|^{-1/2} \phi\left(\Sigma_i^{-1/2}(\mathbf{x} - \mathbf{x}_i)\right),$$

where $\phi(\mathbf{x}) = (2\pi)^{-d/2} \exp(-\mathbf{x}^T\mathbf{x}/2)$ gives the popular Gaussian kernel with covariance matrix Σ_i.

In some cases we may even assume that the kernels have *compact support* properties (see [56] page 153, equations (6.44)) to make them highly localized functions acting in the neighborhood of the observations at which they are anchored.

Remark 3 (Choice of Constraints) Our choice of constraints is guided by the consistency properties of non-parametric estimators. We choose the constraints $\mathscr{C}$ to include the complete sufficient statistic for the unknown process. Without any assumptions the sufficient statistic is simply the whole empirical sample $\mathcal{X}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. Such constraints will hopefully make p a consistent (non-parametric) estimator of the target q^*. One reason for choosing inequality constraints is to make sure that p "dominates" in some sense the unknown q^* by assigning probability mass in the neighborhood of each point $\mathbf{x}_i$ at least as large as the true (estimated) mass. This may make p a good proposal density for an Acceptance Rejection algorithm designed to simulate from q^*. Another reason

for choosing inequality constraints is that they allow us to handle the non-negativity restriction $p(\mathbf{x}) \geqslant 0$ in $\mathscr{P}$ more easily. Moreover, as demonstrated in the examples in the last section, with the inequality constraints the optimal p exhibits model sparsity similar to the sparsity observed with *Support Vector Machines* [66].

Remark 4 (Non-negativity of Density) Note that for some choices of ψ the non-negativity constraint $p(\mathbf{x}) \geqslant 0$ in $\mathscr{P}$ need not be imposed explicitly. We will show that if $\psi(x) = x\ln(x)$, corresponding to the Kullback-Leibler distance, the condition $p(\mathbf{x}) \geqslant 0$ is automatically satisfied. In general however the non-negativity constraint has to be enforced explicitly in the functional optimization.

Remark 5 (Comparison with the CE method) Note that the GCE method solves a **functional** optimization problem to find the optimal posterior density $p(\mathbf{x})$. In contrast the CE method [54] solves the **parametric** optimization problem

$$\min_{\boldsymbol{\theta}} \mathcal{D}\Big(p(\cdot;\boldsymbol{\theta}) \to q^*\Big)$$

to find the optimal CE density $p(\mathbf{x};\boldsymbol{\theta})$ within a pre-specified parametric family $\Big\{p(\cdot;\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta\Big\}$ of densities.

The problem as stated above is a constrained functional optimization problem. More specifically, without the *algebraic* constraint $p(\mathbf{x}) \geqslant 0$, it is an *isoperimetric* Calculus of Variations problem with integral equality and/or inequality constraints (see [50] page 54 and [48]). Since ψ is strictly convex by assumption, the functional (59) is strictly convex and we can use the theory of Lagrangian duality (see [67] pages 219, 266 and 273) to simplify the problem.

3.1 The Dual Optimization Problem

The isoperimetric problem obtained in the previous section is convex and hence there exists a corresponding dual problem. In this case the dual problem is much easier to solve than the primal problem. This is essentially the reason why the strict convexity condition is imposed in the definition of the CE measures. In our case let the **Primal Problem** be:

$$\min_{p} \ \mathcal{D}(p \to q) \tag{62}$$

$$\text{subject to:} \qquad \int p(\mathbf{x})\, K_i(\mathbf{x})\, d\mathbf{x} \ \geqslant \ \kappa_i^*\,, \quad i = 1,\ldots,n \tag{63}$$

$$\int p(\mathbf{x})\, d\mathbf{x} \ = \ 1 \tag{64}$$

Note that the algebraic constraint $p(\mathbf{x}) \geqslant 0,\ \mathbf{x} \in \mathbb{R}^d$ is not included in the formulation of the primal problem. For the time being we assume that the non-negativity constraint is satisfied by the solution of the primal and need not be imposed. To derive the dual corresponding to the primal, define the Lagrangian:

$$\begin{aligned}
& \mathcal{L}(p; \boldsymbol{\lambda}, \lambda_0) = \\
= & \int q(\mathbf{x})\ \psi\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right) d\mathbf{x} - \lambda_0 \left(\int p(\mathbf{x})\ d\mathbf{x} - 1\right) - \sum_{i=1}^{n} \lambda_i \left(\int p(\mathbf{x})\ K_i(\mathbf{x})\ d\mathbf{x} - \kappa_i^*\right) \\
= & \ \lambda_0 + \sum_{i=1}^{n} \lambda_i\ \kappa_i^* + \int \left(q(\mathbf{x})\ \psi\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right) - \lambda_0\ p(\mathbf{x}) - \sum_{i=1}^{n} \lambda_i\ p(\mathbf{x})\ K_i(\mathbf{x})\right) d\mathbf{x} \\
= & \sum_{i=0}^{n} \lambda_i\ \kappa_i^* + \int \left(q(\mathbf{x})\ \psi\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right) - p(\mathbf{x}) \sum_{i=0}^{n} \lambda_i\ K_i(\mathbf{x})\right) d\mathbf{x},
\end{aligned}$$

where for convenience we define $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_n]^T$, $\kappa_0^* = 1$ and $K_0(\cdot) = 1$. Then Calculus of Variations (see [67] page 219) tells us that the **Dual Problem** is:

$$\max_{\boldsymbol{\lambda}, \lambda_0} \ \left\{\inf_p \mathcal{L}(p; \boldsymbol{\lambda}, \lambda_0)\right\} \tag{65}$$

$$\text{subject to:} \qquad \boldsymbol{\lambda} \geqslant \mathbf{0} \tag{66}$$

The dual can be simplified substantially. First $\inf_p \mathcal{L}(p; \boldsymbol{\lambda}, \lambda_0)$ can be calculated explicitly using the Euler-Lagrange equation. In this particular case the Euler-Lagrange equation yields[5]:

$$\psi'\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right) = \sum_{k=0}^{n} \lambda_k\ K_k(\mathbf{x})\ . \tag{67}$$

Since $\psi''(x) > 0$ for $x > 0$, the function $\psi'(x)$ has a unique inverse on the domain $x \in \mathbb{R}^+$. The functional form of the extremal can thus be written explicitly as:

$$p(\mathbf{x}) = q(\mathbf{x})\ {\psi'}^{-1}\left(\sum_{k=0}^{n} \lambda_k\ K_k(\mathbf{x})\right). \tag{68}$$

[5]Since the derivatives of p are not involved the equation is valid in the wider set of PWC functions.

We can then substitute this $p(\mathbf{x})$ into the Lagrangian to obtain:

$$\begin{aligned}\mathcal{L}^*(\boldsymbol{\lambda}, \lambda_0) &= \inf_p \mathcal{L}(p; \boldsymbol{\lambda}, \lambda_0) \\ &= \sum_{i=0}^{n} \lambda_i \, \kappa_i^* + \mathbb{E}_q \, \psi\left(\psi'^{\,-1}\left(\sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{X})\right)\right) \\ &\quad - \sum_{i=0}^{n} \lambda_i \, \mathbb{E}_q \, K_i(\mathbf{X}) \, \psi'^{\,-1}\left(\sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{X})\right).\end{aligned}$$

Then the dual becomes:

$$\max_{\boldsymbol{\lambda}, \lambda_0} \quad \mathcal{L}^*(\boldsymbol{\lambda}, \lambda_0) \,, \tag{69}$$

$$\text{subject to:} \qquad \boldsymbol{\lambda} \geqslant \mathbf{0} \,. \tag{70}$$

Further simplification of $\mathcal{L}^*$ is possible if we set $\Psi' = \psi'^{\,-1}$ and observe that straightforward integration by parts yields:

$$\Psi(x) \;=\; x \, \Psi'(x) - \psi\left(\Psi'(x)\right) + \text{constant} \,.$$

Then $\mathcal{L}^*$ can be written compactly as:

$$\mathcal{L}^*(\boldsymbol{\lambda}, \lambda_0) = \sum_{i=0}^{n} \lambda_i \, \kappa_i^* - \mathbb{E}_q \, \Psi\left(\sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{X})\right), \tag{71}$$

where the constant of integration is ignored as it is irrelevant to the optimization problem. We can finally state the simplest form of the **Dual Problem**:

$$\max_{\boldsymbol{\lambda}, \lambda_0} \quad \sum_{i=0}^{n} \lambda_i \, \kappa_i^* - \mathbb{E}_q \, \Psi\left(\sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{X})\right) \tag{72}$$

$$\text{subject to:} \qquad \boldsymbol{\lambda} \geqslant \mathbf{0}. \tag{73}$$

To get the solution of the **Primal Problem** we apply the transformation:

$$p(\mathbf{x}) = q(\mathbf{x}) \, \Psi'\left(\sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{X})\right). \tag{74}$$

Important quantities for the optimization are the *gradient* and the *Hessian* of $\mathcal{L}^*$:

$$\frac{\partial \mathcal{L}^*}{\partial \lambda_i} \;=\; \kappa_i^* - \mathbb{E}_q \, \Psi'\left(\sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{X})\right) K_i(\mathbf{X}) \tag{75}$$

$$\frac{\partial^2 \mathcal{L}^*}{\partial \lambda_i \partial \lambda_j} = -\,\mathbb{E}_q \, K_i(\mathbf{X}) \, \Psi''\left(\sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{X})\right) K_j(\mathbf{X}) \,, \tag{76}$$

where $i, j \in \{0, 1, \ldots, n\}$. Note that strict concavity of $\mathcal{L}^*$ is equivalent to

$$\sum_{i=0}^{n} \sum_{j=0}^{n} \lambda_i \times \frac{\partial^2 \mathcal{L}^*(\boldsymbol{\lambda}, \lambda_0)}{\partial \lambda_i \partial \lambda_j} \times \lambda_j < 0 \, .$$

This in turn is equivalent to

$$\mathbb{E}_q \left(\sum_{i=0}^{n} \lambda_i \, K_i(\mathbf{X}) \right)^2 \times \Psi'' \left(\sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{X}) \right) > 0 \, ,$$

which is easily shown to be true using $\Psi''(x) = \frac{1}{\psi''\left(\Psi'(x)\right)}$ and (74). This result is in accordance with the general theory of convex optimization (see [8], [1], [7]) which states that if the primal problem is (strictly) convex the dual problem is (strictly) concave and the solution of the primal, which is a (unique) minimizer, coincides exactly with the solution of the dual— a (unique) maximizer. This is usually referred to as *strong duality* (see [7]).

Since there are no constraints on λ_0, the gradient with respect to λ_0 has to be zero:

$$\frac{\partial \mathcal{L}^*}{\partial \lambda_0} = 1 - \mathbb{E}_q \, \Psi' \left(\sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{X}) \right) = 0. \tag{77}$$

$\mathcal{L}^*$ is always a strictly concave function of λ_0 because

$$\frac{\partial^2 \mathcal{L}^*}{\partial \lambda_0^2} = -\mathbb{E}_q \, \Psi'' \left(\sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{X}) \right) < 0 \, .$$

Note that if the *characterizing moment* constraints (61) are strict equalities instead of inequalities then the restriction $\boldsymbol{\lambda} \geqslant \mathbf{0}$ is omitted. Thus with strict equality constraints the dual optimization problem is:

$$\max_{\boldsymbol{\lambda}, \lambda_0} \quad \sum_{i=0}^{n} \lambda_i \, \kappa_i^* - \mathbb{E}_q \, \Psi \left(\sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{X}) \right) , \tag{78}$$

though we may still have to enforce $p(\mathbf{x}) \geqslant 0, \forall \mathbf{x} \in \mathbb{R}^d$ explicitly. Special cases of (78) are:

The MCE algorithm [53]

Choose $\psi(x) = x \ln(x) - x$, then $\psi'^{-1}(x) = \exp(x) = \Psi'(x) = \Psi(x)$, $p^*(\mathbf{x}) = q(\mathbf{x}) \exp\left(\sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{x})\right) \geqslant 0$ and $\mathcal{D}(p \to q) = \int p(\mathbf{x}) \ln\left(p(\mathbf{x})/q(\mathbf{x})\right) d\mathbf{x} - 1$. The Lagrange multipliers are determined from the maximization of the dual (78). In

this case there are no constraints on $\boldsymbol{\lambda}$ and λ_0, the unconstrained maximization of the strictly concave $\mathcal{L}^*$ leads to the set of non-linear equations for $\nabla_\lambda \mathcal{L}^* = \mathbf{0}$:

$$\mathbb{E}_q \exp\left(\sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{X})\right) K_i(\mathbf{X}) = \kappa_i^* \,, \quad i = 0, \ldots, n \tag{79}$$

The solution gives the unique optimal $p(\mathbf{x})$ for the MCE method. We thus make the conclusion that the MCE method is equivalent to choosing the proposal density[6]

$$p(\mathbf{x}) = \frac{q(\mathbf{x}) \exp\left(\sum_{k=1}^{n} \lambda_k \, K_k(\mathbf{x})\right)}{\mathbb{E}_q \exp\left(\sum_{k=1}^{n} \lambda_k \, K_k(\mathbf{X})\right)} \tag{80}$$

from the General Exponential Family [49] and then minimizing what appears to be a distance measure:

$$\min_{\lambda} \quad -\mathcal{L}^*(\lambda) = \mathbb{E}_{q^*} \ln \frac{q(\mathbf{X})}{p(\mathbf{X})}$$

without any constraints on the multipliers. An advantage of the MCE method is that $\kappa_i^* = \frac{1}{J} \sum_{j=1}^{J} K_i(\mathbf{X}_j)$, with $\mathbf{X}_1, \ldots, \mathbf{X}_J \sim q^*$, is the asymptotically efficient (i.e. Maximum likelihood) estimator of $\mathbb{E}_{q^*} K_i(\mathbf{X})$. This is a consequence of the fact that p in this case belongs to the General Exponential Family of probability functions. The salient features of the MCE method can be summarized as follows:

1. In the MCE method the dual (78) of the primal functional optimization problem becomes a GPP.

2. The expectations/integrals on the left-hand side of (79) have to be estimated via an empirical average to give the *stochastic counterpart* of (79):

$$\frac{1}{n} \sum_{j=1}^{n} \exp\left(\sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{X}_j)\right) K_i(\mathbf{X}_j) = \kappa_i^*, \quad \{\mathbf{X}_j\}_{j=1}^{n} \sim q, \quad i = 0, \ldots, n.$$

3. Simulation from (80) and any other member of the General Exponential Family is in general feasible only via the Accept-Reject algorithm.

4. The non-negativity of (80) is ensured by its exponential functional form. This makes the optimization easier.

5. The MCE optimal density (80) does not conform to the functional form of the asymptotically consistent density estimator (52) in the General Kernel Theorem. If q^* does not belong to the General Exponential Family, then the MCE optimal density (80) may not converge to q^* as $n \to \infty$.

[6] Note that we have substituted for λ_0.

6. While the functional form of (80) is not asymptotically optimal, the estimation of the characterizing moments $\mathbb{E}_{q^*} K_i(\mathbf{X})$ through $\kappa_i^* = \sum_{j=1}^J K_i(\mathbf{X}_j)$, $\mathbf{X}_1, \ldots, \mathbf{X}_J \sim q^*$, is asymptotically optimal.

The CE algorithm [54]

If in the CE method we choose a sampling density from the General Exponential Family $p(\mathbf{x}) = \frac{\exp\left(\sum_{k=1}^n \lambda_k K_k(\mathbf{x})\right)}{\int \exp\left(\sum_{k=1}^n \lambda_k K_k(\mathbf{x})\right) d\mathbf{x}}$, then Maximizing the Likelihood [7] $\sum_{j=1}^J \ln p(\mathbf{X}_j)$, where $\mathbf{X}_1, \ldots, \mathbf{X}_J \sim q^*$, gives the CE updating equations ($i = 0, \ldots, n$):

$$\frac{\int \exp\left(\sum_{k=1}^n \lambda_k K_k(\mathbf{x})\right) K_i(\mathbf{x})\, d\mathbf{x}}{\int \exp\left(\sum_{k=1}^n \lambda_k K_k(\mathbf{x})\right) d\mathbf{x}} = \frac{1}{J} \sum_{j=1}^J K_i(\mathbf{X}_j) = \kappa_i^*, \quad \{\mathbf{X}_j\}_{j=1}^J \sim q^*$$

for the parameters $\{\lambda_i\}_{i=0}^n$. We thus conclude that the updating rules of the CE method (see [54] pages 68, 69 and Example 3.5) coincide with the updating rules of the GCE method in cases where

1. the CE method chooses a sampling/proposal density p from the General Exponential Family with *natural parameters* $\{\lambda_k\}_{k=1}^n$ and *natural statistics* $\{K_k(\mathbf{x})\}_{k=1}^n$ (see [49] page 95) and

2. the GCE method chooses the convex $\Psi(x) = \exp(x)$ in (78).

The updating rules between the two methods do not agree under any other conditions. Again note that the Maximum Likelihood estimators of parameters of densities in the General Exponential Family achieve the so called Cramer-Rao lower bound (see [49] page 223). This makes the simple estimator $\kappa_i^* = \frac{1}{J} \sum_{j=1}^J K_i(\mathbf{X}_j)$, $\{\mathbf{X}_j\}_{j=1}^J \sim q^*$ the MVUE of $\mathbb{E}_{q^*} K_i(\mathbf{X})$. This is the advantage of using a proposal density from the General Exponential Family. Note, however, that typically we have random variables from the prior q and not from the target q^*. In this case the CE method uses the Likelihood Ratio (LR) estimator

$$\kappa_i^* = \frac{\sum_{j=1}^J W(\mathbf{X}_j)\, K_i(\mathbf{X}_j)}{\sum_{j=1}^J W(\mathbf{X}_j)}, \quad W(\mathbf{X}_j) = \frac{q^*(\mathbf{X}_j)}{q(\mathbf{X}_j)}, \quad \{\mathbf{X}_j\}_{j=1}^J \sim q. \tag{81}$$

Since (81) no longer follows from the Maximum Likelihood Principle [49], the optimality of the LR estimator (81) is dubious and still an important problem of research (see [18]).

[7] maximizing the Likelihood is the same as minimizing Burg's CE distance $\mathbb{E}_{q^*} \ln(q^*(\mathbf{X})/p(\mathbf{X}))$ between q^* and p. Minimization of Burg's CE distance is the highlighting feature of the CE method.

3.2 The choice for ψ

Our present aim is to choose the function ψ in Csiszár's measure such that:

1. The integral/expectation in (72) can be done analytically or at least without too much trouble.

2. Maximizing (72)+(73) and hence finding the set of Lagrange multipliers $\{\lambda_k\}_{k=0}^n$ is relatively easy. E.g., if ${\psi'}^{-1} = \Psi'$ are linear then (75) is linear in the Lagrange multipliers and the Hessian matrix (76) is constant. This can greatly simplify the optimization.

3. Generating random variates from the extremal pdf p in (68) is relatively easy. E.g., if Ψ' is linear then p is a discrete mixture and the *composition method* (also known as the *convolution method*) for random variate generation applies.

Satisfying these desiderata simultaneously is only possible for few specific choices of ψ . In particular we can choose Ψ' to be linear. Then ψ' is linear and the definition of Csiszár's measure requires:

$$\begin{aligned} \psi'(x) &= ax + b \\ \psi''(x) &> 0 \\ \psi(1) &= 0, \end{aligned}$$

hence :

$$\psi(x) = \frac{a}{2}(x^2 - 1) + b(x - 1), \quad a > 0$$

for arbitrary constants $a > 0$ and b . Then Csiszár's measure can be written as:

$$\begin{aligned} \mathcal{D}(p \to q) &= \frac{a}{2} \int q(\mathbf{x}) \left(\frac{p^2(\mathbf{x})}{q^2(\mathbf{x})} - 1 \right) d\mathbf{x} \\ &= -\frac{a}{2} + \frac{a}{2} \int \frac{p^2(\mathbf{x})}{q(\mathbf{x})} \, d\mathbf{x} \\ &= \frac{a}{2} \int \frac{\big(p(\mathbf{x}) - q(\mathbf{x})\big)^2}{q(\mathbf{x})} \, d\mathbf{x}. \end{aligned}$$

Note that for optimization purposes the value of a is irrelevant as long as $a > 0$. We will thus choose $a = \frac{1}{2}$ to obtain:

$$\mathcal{D}_2(p \to q) = \frac{1}{2} \int \left(\frac{p^2(\mathbf{x})}{q(\mathbf{x})} - q(\mathbf{x}) \right) d\mathbf{x},$$

which is Pearson's χ^2 CE distance. The choice $\psi(x) = \frac{1}{2}(x^2 - 1)$ (note that the linear term $b(x - 1)$ is irrelevant and hence is omitted) ensures that:

1. ${\psi'}^{-1}(x) = x = \Psi'(x)$ allowing us to write (72) as a linear combination of integrals/expectations each of which, for various kernel functions K_i, can be evaluated analytically.

2. The Hessian matrix (76) of (72) is independent of the Lagrange multipliers.

3. The resulting density function (68) can be simulated using the *composition method*.

In fact (68) becomes the *particle filter* density[8]:

$$p(\mathbf{x}) = q(\mathbf{x}) \sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{x}) \tag{82}$$

and the dual problem (72)+(73) becomes:

$$\max_{\boldsymbol{\lambda},\lambda_0} \quad -\frac{1}{2} + \sum_{i=0}^{n} \lambda_i \, \kappa_i^* - \frac{1}{2} \sum_{i=0}^{n} \sum_{j=0}^{n} \lambda_i \, \lambda_j \, \mathbb{E}_q K_i(\mathbf{X}) K_j(\mathbf{X}), \tag{83}$$

$$\text{subject to:} \quad \boldsymbol{\lambda} \geqslant \mathbf{0}. \tag{84}$$

This optimization is equivalent to :

$$\min_{\boldsymbol{\lambda},\lambda_0} \quad \frac{1}{2} \int \frac{\left(p(\mathbf{x}) - q^*(\mathbf{x})\right)^2}{q(\mathbf{x})} \, d\mathbf{x}, \tag{85}$$

$$\text{subject to:} \quad \boldsymbol{\lambda} \geqslant \mathbf{0}, \tag{86}$$

with $p(\mathbf{x}) = q(\mathbf{x}) \sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{x})$. Thus this approach is equivalent to choosing a discrete mixture of kernel functions as the sampling density and then minimizing the *projection pursuit index* (85) (see [62], page 129) between the sampling and the target density q^*. We now proceed to rewrite the dual problem in a form which is easier to interpret. First since there are no constraints on λ_0 we can solve $\frac{\partial \mathcal{L}^*}{\partial \lambda_0} = 0$ in (75) and determine λ_0 as a function of $\boldsymbol{\lambda}$:

$$\lambda_0 = 1 - \mathbb{E}_q \sum_{k=1}^{n} \lambda_k \, K_k(\mathbf{X}) = 1 - \sum_{k=1}^{n} \lambda_k \, \mathbb{E}_q K_k(\mathbf{X}).$$

We then substitute for λ_0 to obtain

$$p(\mathbf{x}) = q(\mathbf{x}) + q(\mathbf{x}) \sum_{k=1}^{n} \lambda_k \left(K_k(\mathbf{x}) - \mathbb{E}_q K_k(\mathbf{X}) \right). \tag{87}$$

[8] In the particle filter context (see [18]) the set $\{\lambda_i\}_{i=0}^n$ is the set of Sampling Importance Resampling (SIR) weights.

The Lagrange multipliers are determined from optimization of the dual:

$$\max_{\boldsymbol{\lambda}} \sum_{i=1}^{n} \lambda_i \left(\kappa_i^* - \mathbb{E}_q K_i(\mathbf{X})\right) - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \, \lambda_j \, \mathrm{Cov}_q \left(K_i(\mathbf{X}); K_j(\mathbf{X})\right),$$

subject to $\boldsymbol{\lambda} \geqslant \mathbf{0}$. The quadratic form of the problem can be written in matrix notation:

$$\max_{\boldsymbol{\lambda}} \quad -\frac{1}{2}\boldsymbol{\lambda}^T C \boldsymbol{\lambda} + \mathbf{c}^T \boldsymbol{\lambda} \tag{88}$$

$$\text{subject to:} \quad \boldsymbol{\lambda} \geqslant \mathbf{0}, \tag{89}$$

where

$$\begin{aligned} C &= \mathbb{E}_q \left[\Big(\mathbf{K}(\mathbf{X}) - \boldsymbol{\kappa}\Big)\Big(\mathbf{K}(\mathbf{X}) - \boldsymbol{\kappa}\Big)^T \right] \\ \mathbf{c} &= \boldsymbol{\kappa}^* - \boldsymbol{\kappa} \\ \text{with} \qquad \mathbf{K}(\mathbf{x}) &= \Big[K_1(\mathbf{x}) \, K_2(\mathbf{x}) \cdots K_n(\mathbf{x})\Big]^T \\ \boldsymbol{\kappa} &= \mathbb{E}_q \, \mathbf{K}(\mathbf{X}) \\ \boldsymbol{\kappa}^* &= [\ \kappa_1^* \ \ \kappa_2^* \ \ \kappa_3^* \ \ldots \ \kappa_{n-1}^* \ \ \kappa_n^* \ ]^T. \end{aligned}$$

Choosing $\psi(x) = \frac{1}{2}(x^2 - 1)$ thus makes the optimization problem (72)+(73) a Quadratic Programming Problem (QPP) for the Lagrange multipliers. For theoretical and computational convenience we now rescale the QPP. Let $V = \mathrm{diag}\big(C\big)$ and $\boldsymbol{\nu} = V^{1/2}\boldsymbol{\lambda}$. Then $C = V^{1/2} A V^{1/2}$, where A is the correlation matrix corresponding to the covariance matrix C, and (88)+(89) is equivalent to:

$$\max_{\boldsymbol{\nu}} \quad -\frac{1}{2}\boldsymbol{\nu}^T A \boldsymbol{\nu} + \mathbf{a}^T \boldsymbol{\nu} \tag{90}$$

$$\text{subject to:} \quad \boldsymbol{\nu} \geqslant \mathbf{0}, \tag{91}$$

where

$$\boldsymbol{\lambda} = V^{-1/2}\boldsymbol{\nu} \tag{92}$$

$$\mathbf{a} = V^{-1/2}\mathbf{c} = V^{-1/2}\Big(\boldsymbol{\kappa}^* - \boldsymbol{\kappa}\Big) \tag{93}$$

$$A = V^{-1/2} C V^{-1/2} \equiv \Big[\mathrm{Corr}_q \left(K_i(\mathbf{X}); K_j(\mathbf{X})\right) \Big]_{ij} \tag{94}$$

The two problems (90)+(91) and (88)+(89) are equivalent theoretically but not numerically. The problem (90)+(91) is intuitively easier to understand and numerically better behaved, because the entries of A are always bounded between -1 and 1. Although A may be numerically ill-conditioned, with probability one A is a positive definite symmetric correlation matrix. Therefore

any of the QPP (90),(88) and (83) are strictly convex and the KKT conditions guarantee a unique global minimum for any concave constraints. In particular (90),(88) and (83) have a unique global minimum under the concave constraints (91), (89). The solution (c.f. (82)) in matrix form is:

$$p(\mathbf{x}) = q(\mathbf{x}) \left(\lambda_0 + \boldsymbol{\lambda}^T \mathbf{K}(\mathbf{x})\right), \tag{95}$$

$$\text{where} \quad \lambda_0 = 1 - \boldsymbol{\kappa}^T \boldsymbol{\lambda} \tag{96}$$

$$= 1 - \boldsymbol{\kappa}^T V^{-1/2} \boldsymbol{\nu}\,. \tag{97}$$

However the solution of the QPP may not be useful because:

1. For a negative λ_0, (95) may take negative values for some $\mathbf{x}$. This is unacceptable for a probability density function.

2. Even if $p(\mathbf{x}) \geqslant 0$ for all $\mathbf{x} \in \mathbb{R}^d$, a negative λ_0 makes (95) a mixture density with a negative weight. Sampling from it will require the use of the Accept-Reject algorithm, which can be highly inefficient in high dimensions.

We now explore under what conditions the above problems are avoided. It turns out that we can always rescale the kernels $\{K_k\}_{k=1}^n$ so that $\lambda_0 \geqslant 0$. More precisely we can find a set of bandwidth parameters $\{\Sigma_i\}_{i=1}^n \in \mathscr{S}\left(\{\Sigma_i\}_{i=1}^n\right)$, where

$$\mathscr{S}\left(\{\Sigma_i\}_{i=1}^n\right) = \left\{\{\Sigma_i\}_{i=1}^n : \ \boldsymbol{\kappa}^T \boldsymbol{\lambda}^* \leqslant 1, \ \boldsymbol{\lambda}^* = \underset{\boldsymbol{\lambda} \geqslant \mathbf{0}}{\operatorname{argmin}} \left[\boldsymbol{\lambda}^T C \boldsymbol{\lambda} - 2\, \boldsymbol{\lambda}^T \mathbf{c}\right]\right\}$$

and C, $\mathbf{c}$ and $\boldsymbol{\kappa}$ depend implicitly on the argument of $\mathscr{S}$ through the kernels $\{K(\mathbf{x}; \mathbf{x}_i, \Sigma_i)\}_{i=1}^n$. The set $\mathscr{S}$ is the set of *admissible* bandwidth parameters in the sense that

$$\{\Sigma_i\}_{i=1}^n \in \mathscr{S}\left(\{\Sigma_i\}_{i=1}^n\right) \Leftrightarrow p \in \mathscr{P}.$$

If for simplicity we have a single bandwidth $\Sigma_i = \sigma I, \ \forall i$ for all the kernels $\{K_i(\mathbf{x})\}_{i=1}^n = \{K(\mathbf{x}; \mathbf{x}_i, \sigma I)\}_{i=1}^n$, then the solution of the dual QPP with $\sigma I \in \mathscr{S}(\sigma I)$ ensures that the solution of the primal $p \in \mathscr{P}$.

Remark 6 There are two extreme values for σI. We may either choose σI such that $\lambda_0 = 1$, in which case we assign maximum weight to the prior q, or we can choose σI such that $\lambda_0 = 0$, in which case we eliminate the prior as a mixture component in (95). Values for σI in between these two extremes represent a trade-off between the prior density and the observed empirical data. Note that $\lambda_0 = 1$ is equivalent to $\boldsymbol{\lambda} = \mathbf{0}$ which is only possible if q is such that $\boldsymbol{\kappa} \geqslant \boldsymbol{\kappa}^*$. So it may not be always possible to assign all the probability mass to the prior q and obtain $p(\mathbf{x}) = q(\mathbf{x})$ for (95).

Remark 7 If q is an improper prior then we must choose σI such that $\lambda_0 = 0$. This is the only choice which will guarantee that (95) is a proper pdf and not a mixture pdf with component q. Usually choosing σI such that $\lambda_0 = 0$ gives a unique value for the bandwidth σI.

Remark 8 (Pointwise non-negativity constraint) If $\sigma I \notin \mathscr{S}(\sigma I)$, then the only other way to make sure that $p \in \mathscr{P}$ is to solve the primal problem with the addition of the pointwise inequality constraint $p(\mathbf{x}) \geqslant 0$. From the preliminary section we know the solution has the form:

$$\check{p}(\mathbf{x}) \;=\; \begin{cases} p(\mathbf{x}), & \mathbf{x} \in \mathcal{S} \\ 0, & \mathbf{x} \notin \mathcal{S} \end{cases}, \tag{98}$$

where $p(\mathbf{x})$ is the extremal of the primal problem (without the pointwise constraint), $p(\mathbf{x}) \geqslant 0, \forall \mathbf{x} \in \mathcal{S}$ and the boundary of the set $\mathcal{S}$ is determined from the multidimensional analogue of the transversality and continuity condition. The addition of the pointwise constraint makes the primal problem a computationally difficult Calculus of Variations problem comparable to solving a multidimensional Differential equation. Essentially finding $\check{p}$ involves first identifying the set $\mathcal{S}$ for which the solution of the primal $p(\mathbf{x}) \geqslant 0$ and then resolving the primal over this set (taking all the integrals in the definition of the primal over $\mathcal{S} \subset \mathscr{X}$). Identifying the set $\mathcal{S}$ is an infinite dimensional problem and there is no duality trick which can get around the problem. Moreover assuming that we can somehow obtain $\check{p}$, sampling from $\check{p}$ will only be possible with the Accept-Reject method. This is undesirable, since in keeping with the curse of dimensionality, the efficiency of the Accept-Reject method decays exponentially as the dimension of $\mathscr{X}$ increases.

In summary $\{\Sigma_i\}_{i=1}^n \in \mathscr{S}(\{\Sigma_i\}_{i=1}^n)$ implies that:

1. The solution (95) belongs to $\mathscr{P}$, i.e., is non-negative and integrates to one.
2. Simulation from (95) via the composition method is relatively easy. Thus $p(k) = \int p(\mathbf{x}, k)\, d\mathbf{x}$, where the joint density

$$p(\mathbf{x}, k) = \begin{cases} \lambda_0\, q(\mathbf{x}), & \text{for} \quad k = 0 \\ \lambda_k\, q(\mathbf{x})\, K_k(\mathbf{x}), & \text{for} \quad k = 1, \ldots, n \end{cases},$$

   is a proper pmf. of the index k.

This is explained in greater detail below.

3.3 Sampling from p

Define

$$\begin{aligned}\mathbf{w} = [w_0, w_1, \ldots, w_n]^T &= \begin{bmatrix} \lambda_0 \\ \text{diag}(\boldsymbol{\kappa})\ \boldsymbol{\lambda} \end{bmatrix} \\ &= \begin{bmatrix} 1 - \boldsymbol{\kappa}^T V^{-1/2}\ \boldsymbol{\nu} \\ \text{diag}(\boldsymbol{\kappa})\ V^{-1/2}\boldsymbol{\nu} \end{bmatrix},\end{aligned}$$

then the distribution of the index k is:

$$p(k) = w_k = \begin{cases} \lambda_0 = 1 - \boldsymbol{\kappa}^T\boldsymbol{\lambda} = 1 - \boldsymbol{\kappa}^T V^{-1/2}\ \boldsymbol{\nu}, & k = 0 \\ \lambda_k\ \kappa_k = \nu_k \frac{\mathbb{E}_q K_k(\mathbf{X})}{\sqrt{\text{Var}_q K_k(\mathbf{X})}}, & k = 1, \ldots, n \end{cases}$$

and

$$p(\mathbf{x}\,|\,k) = \frac{q(\mathbf{x})K_k(\mathbf{x})}{\mathbb{E}_q K_k(\mathbf{X})} = \frac{q(\mathbf{x})K_k(\mathbf{x})}{\kappa_k}, \qquad k = 0, \ldots, n.$$

So simulation from

$$p(\mathbf{x}) = \mathbf{w}^T \begin{bmatrix} q(\mathbf{x}) \\ \text{diag}(\boldsymbol{\kappa})^{-1}\mathbf{K}(\mathbf{x})\ q(\mathbf{x}) \end{bmatrix} = \sum_{k=0}^{n} w_k\ \frac{q(\mathbf{x})K_k(\mathbf{x})}{\mathbb{E}_q K_k(\mathbf{X})}$$

is accomplished by sampling from the joint density

$$\{\mathbf{X}_j, \mathrm{K}_j\}_{j=1}^{J} \overset{i.i.d}{\sim} p(\mathbf{x}, k) = p(k) \times p(\mathbf{x}\,|\,k) = w_k \times \frac{q(\mathbf{x})K_k(\mathbf{x})}{\mathbb{E}_q K_k(\mathbf{X})}$$

using the stratification algorithm outlined in the preliminary section. We do not discard the index variables $\{\mathrm{K}_j\}_{j=1}^{J}$ as they carry useful information and can be used to simplify various calculations on the next iteration.

3.4 Choosing K

For many choices of the kernel functions $\mathbf{K}$ we can find $\text{Corr}_q\left(K_i(\mathbf{X});\ K_j(\mathbf{X})\right)$ or $\text{Cov}_q\left(K_i(\mathbf{X});\ K_j(\mathbf{X})\right)$ analytically, provided q itself is another linear combination of kernels or a uniform prior. In practice the choice for K is dictated by the assumptions we make about the smoothness of the target density q^*. If q^* is known to be smooth then K should also be smooth. Naturally the more smooth q^* is, the easier it is to estimate. We now give two examples of possible kernels. The calculations are long but straightforward and only the final results are presented. The purpose is to show that only $\boldsymbol{\kappa}^*$ needs to be estimated via a Monte Carlo sample and all the other elements of the QPP can be calculated analytically for a wide variety of kernels.

Example 10 (Uniform kernel) If we have no prior smoothness information about q^* then we choose the uniform kernel:

$$K_i(\mathbf{x}) = K(\mathbf{x}; \mathbf{x}_i, \Sigma) = \prod_{l=1}^{d} \mathcal{K}(\mathbf{x}(l); \mathbf{x}_i(l), \sigma(l)), \tag{99}$$

$$\text{where} \quad \mathcal{K}(x; x_i, \sigma) = \frac{I\{|x - x_i| < \sigma/2\}}{\sigma} \tag{100}$$

and:

1. For simplicity $\Sigma_i = \Sigma = \text{diag}(\boldsymbol{\sigma}),\ \forall i$ is assumed to be a diagonal matrix providing different smoothing for each of the d dimensions.
2. $\mathcal{K}$ is a univariate kernel function. A multivariate kernel K can in general be constructed as the product of univariate kernels.

Given our complete ignorance about q^*, we choose as prior $q \propto 1$ on $\mathbb{R}^d$. After some straightforward calculations:

$$\begin{aligned}
\sigma \int_{-\infty}^{\infty} \mathcal{K}(x; x_i, \sigma)\, \mathcal{K}(x; x_j, \sigma)\, dx &= \left(1 - \frac{|x_i - x_j|}{\sigma}\right) I\left\{|x_i - x_j| < \sigma\right\}, \\
\prod_{l=1}^{d} \sigma(l) \int_{\mathbb{R}^d} K_i(\mathbf{x})\, K_j(\mathbf{x})\, d\mathbf{x} &= I\left\{|\mathbf{x}_i - \mathbf{x}_j| < \boldsymbol{\sigma}\right\} \prod_{l=1}^{d} \left(1 - \frac{\left|\mathbf{x}_i(l) - \mathbf{x}_j(l)\right|}{\boldsymbol{\sigma}(l)}\right), \\
\prod_{l=1}^{d} \sigma(l)\, \kappa_i^* &= \frac{1}{n} \sum_{k=1}^{n} I\{|\mathbf{X}_k - \mathbf{x}_i| < \boldsymbol{\sigma}/2\}, \quad \mathbf{X}_k \sim q^*,
\end{aligned}$$

where κ_i^* is estimated using a sample from q^*. The uniform kernel has the advantage of computational simplicity due to its highly localized nature. E.g., the problem (83)+(84) is a very sparse QPP because $\int_{\mathbb{R}^d} K_i(\mathbf{x})\, K_j(\mathbf{x})\, d\mathbf{x}$ is zero for distant $\mathbf{x}_i$ and $\mathbf{x}_j$. This sparsity is observed to speed up the solution of the QPP dramatically and can be useful for problems with large sample $\mathcal{X}_n$.

Example 11 (Gaussian kernel) Suppose that $\phi(\mathbf{x}) = (2\pi)^{-d/2} \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{x}\right)$ and we choose a Gaussian kernel

$$K_i(\mathbf{x}) = K(\mathbf{x}; \mathbf{x}_i, \Sigma_i) = |\Sigma_i|^{-1/2} \phi\left(\Sigma_i^{-1/2}(\mathbf{x} - \mathbf{x}_i)\right).$$

In other words $K_i(\mathbf{x}; \mathbf{x}_i, \Sigma_i)$ is the multivariate normal density $\mathsf{N}(\mathbf{x}_i, \Sigma_i)$. Assume that $q(\mathbf{x}) \propto 1$ for all $\mathbf{x} \in \mathscr{X} \equiv \mathbb{R}^d$, i.e., q is the improper uniform prior. Then using the results in the preliminary section the QPP (83)+(84) is simplified using:

$$\begin{aligned}
\int_{\mathbb{R}^d} K_i(\mathbf{x})\, K_j(\mathbf{x})\, d\mathbf{x} &= |\Sigma_i + \Sigma_j|^{-1/2} \phi\left(\left(\Sigma_i + \Sigma_j\right)^{-1/2} \left(\mathbf{x}_i - \mathbf{x}_j\right)\right) \\
\kappa_i^* &= \frac{1}{n} \sum_{k=1}^{n} |\Sigma_i|^{-1/2} \phi\left(\Sigma_i^{-1/2}(\mathbf{X}_k - \mathbf{x}_i)\right), \quad \mathbf{X}_k \sim q^*.
\end{aligned}$$

There may be some problems when using the same sample points $\mathcal{X}_n$ as both location parameters for the kernels and as a sample in the estimation of $\boldsymbol{\kappa}^*$. This problem is addressed next.

3.5 Estimating $\boldsymbol{\kappa}^*$

Assume we use the same set $\mathcal{X}_n = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ as both kernel location parameters and as a sample for the estimation of $\boldsymbol{\kappa}^*$. The simplest unbiased estimator of $\mathbb{E}_{q^*} K_i(\mathbf{X})$ is:

$$\kappa_i^* = \frac{1}{n-1} \sum_{j \neq i}^{n} K_i(\mathbf{X}_j),$$

where we have assumed $\mathbf{X}_1, \ldots, \mathbf{X}_n \sim q^*$. This is the *cross-validatory*, also known as *leave-one-out*, estimator and its consistency properties are established in [9], [10], [2] and [63]. The simplest argument against the inclusion of the $i-$th observation in the estimate is the following. K_i is anchored at the i-th observation and we wish to estimate the probability mass which q^* assigns in the neighborhood of the $i-$th observation. For a given fixed anchor point $\mathbf{x}_i$ the probability that a random draw from q^* equals $\mathbf{x}_i$ is zero. Hence we should not use $\mathbf{x}_i$ both as an anchor point and as a random draw from q^*.

We assumed that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ have density q^*. This is rarely possible since q^* is very complicated. Instead the data are typically drawn from a proposal density— the prior q. In such cases we use the unbiased importance sampling (IS) estimator:

$$\kappa_i^* = \sum_{j \neq i} \frac{q^*(\mathbf{X}_j)}{q(\mathbf{X}_j)} K_i(\mathbf{X}_j), \quad \mathbf{X}_1, \ldots, \mathbf{X}_n \sim q,$$

where a standard approach (see [18]) is to normalize the IS weights $\left\{ \frac{q^*(\mathbf{X}_j)}{q(\mathbf{X}_j)} \right\}_{i=1}^{n}$ in the hope of reducing the variance of the estimator. Moreover if $q(\mathbf{x}) = \sum_k q(k)\, q(\mathbf{x} \,|\, k)$ is a discrete mixture then we can use the unbiased estimator:

$$\kappa_i^* = \frac{1}{n-1} \sum_{j \neq i} \frac{q^*(\mathbf{X}_j)}{q(\mathbf{X}_j \,|\, \mathrm{K}_j)} K_i(\mathbf{X}_j), \quad \{\mathbf{X}_j, \mathrm{K}_j\}_{j=1}^{J} \overset{i.i.d}{\sim} q(\mathbf{x}, k).$$

This estimator is computationally efficient because $q(\mathbf{x} \,|\, k)$ is cheaper to evaluate than $q(\mathbf{x})$. This is the reason why we keep the index set $\{K_i\}_{i=1}^{n}$ used to generate random variables from the kernel pdfs at each iteration of the learning algorithm.

Example 12 (Minimum Variance IS Density) Suppose we use the set of Gaussian kernels

$$K_i(\mathbf{x}) = |\Sigma_i|^{-1/2} \times \phi\left(\Sigma_i^{-1/2} (\mathbf{x} - \mathbf{x}_i)\right), \quad i = 1, \ldots, n$$

and the target density is

$$q^*(\mathbf{x}) = \frac{I\{S(\mathbf{x}) > \gamma\} f(\mathbf{x})}{\ell} = \frac{\varphi_\gamma(\mathbf{x})}{\ell},$$

i.e., the minimum variance IS density [54] for the estimation of $\ell = \mathbb{E}_f I\{S(\mathbf{X}) > \gamma\} = \mathbb{P}_f(S(\mathbf{X}) > \gamma)$. Suppose further that the prior is a mixture of Gaussians:

$$q(\mathbf{x}) = \sum_k \omega_k \, |\Lambda_k|^{-1/2} \phi\left(\Lambda_k^{-1/2}\left(\mathbf{x} - \boldsymbol{\mu}_k\right)\right) = \sum_k q(k)\, q(\mathbf{x} \,|\, k) = \sum_k q(\mathbf{x}; k).$$

Then

$$\kappa_i^* = \sum_{j \neq i} \frac{\varphi_\gamma(\mathbf{X}_j)}{\hat{\ell}} \frac{|\Lambda_{\mathrm{K}_j}|^{1/2}}{|\Sigma_j|^{1/2}} \exp\left(\frac{1}{2}\left(\mathbf{X}_j - \boldsymbol{\mu}_{\mathrm{K}_j}\right)^T \Lambda_{\mathrm{K}_j}^{-1}\left(\mathbf{X}_j - \boldsymbol{\mu}_{\mathrm{K}_j}\right) - \frac{1}{2}\left(\mathbf{X}_j - \mathbf{x}_i\right)^T \Sigma_i^{-1}\left(\mathbf{X}_j - \mathbf{x}_i\right)\right),$$

where

$$\left\{\mathbf{X}_j, \mathrm{K}_j\right\}_{j=1}^n \overset{i.i.d}{\sim} q(\mathbf{x}, k) \equiv \omega_k \times |\Lambda_k|^{-1/2} \phi\left(\Lambda_k^{-1/2}\left(\mathbf{x} - \boldsymbol{\mu}_k\right)\right)$$

and

$$\hat{\ell} = (2\pi)^{d/2} \sum_j \varphi_\gamma(\mathbf{X}_j) \, |\Lambda_{\mathrm{K}_j}|^{1/2} \exp\left(\frac{1}{2}\left(\mathbf{X}_j - \boldsymbol{\mu}_{\mathrm{K}_j}\right)^T \Lambda_{\mathrm{K}_j}^{-1}\left(\mathbf{X}_j - \boldsymbol{\mu}_{\mathrm{K}_j}\right)\right).$$

Example 13 (Boltzmann Density) Suppose that the target density is

$$q^*(\mathbf{x}) = \frac{\mathrm{e}^{-\gamma S(\mathbf{x})}}{\ell}, \quad \gamma > 0, \quad S : \mathbb{R}^d \to \mathbb{R}^+.$$

Then:

$$\kappa_i^* = \sum_{j \neq i} \frac{\left|\Lambda_{\mathrm{K}_j}\right|^{1/2}}{\hat{\ell}\left|\Sigma_j\right|^{1/2}} \exp\left(\frac{1}{2}\left(\mathbf{X}_j - \boldsymbol{\mu}_{\mathrm{K}_j}\right)^T \Lambda_{\mathrm{K}_j}^{-1}\left(\mathbf{X}_j - \boldsymbol{\mu}_{\mathrm{K}_j}\right) - \frac{1}{2}\left(\mathbf{X}_j - \mathbf{x}_i\right)^T \Sigma_i^{-1}\left(\mathbf{X}_j - \mathbf{x}_i\right) - \gamma S(\mathbf{X}_j)\right),$$

where

$$\left\{\mathbf{X}_j, \mathrm{K}_j\right\}_{j=1}^n \overset{i.i.d}{\sim} q(\mathbf{x}; k)$$

and

$$\hat{\ell} = (2\pi)^{d/2} \sum_j \left|\Lambda_{\mathrm{K}_j}\right|^{1/2} \exp\left(\frac{1}{2}\left(\mathbf{X}_j - \boldsymbol{\mu}_{\mathrm{K}_j}\right)^T \Lambda_{\mathrm{K}_j}^{-1}\left(\mathbf{X}_j - \boldsymbol{\mu}_{\mathrm{K}_j}\right) - \gamma S(\mathbf{X}_j)\right).$$

3.6 Choosing $\{\Sigma_i\}_{i=1}^n$

In general, in order of increasing complexity, we can consider any of these choices for the bandwidth matrices:

1. $\Sigma_i = \sigma \operatorname{diag}(\mathbf{1}) = \sigma I, \ \forall i$, then we simply choose $\sigma I \in \mathscr{S}(\sigma I)$ and the problem is solved. This procedure usually gives a unique bandwidth.

2. $\Sigma_i = \hat{\sigma}_i I h$, where h is a common scale parameter , then an asymptotically justified procedure due to Abramson (see [3], [56], [61]) is to construct a rough pilot estimate $\hat{q}^*$ of q^* and take $\hat{\sigma}_i \propto [q^*(\mathbf{x}_i)]^{-1/2}, \ \forall i$. We then find a global scale parameter h such that $\{\hat{\sigma}_i I h\}_{i=1}^n \in \mathscr{S}\left(\{\hat{\sigma}_i I h\}_{i=1}^n\right)$, i.e., $p \in \mathscr{P}$.

3. If $\Sigma_i = h \ \operatorname{diag}(\boldsymbol{\sigma}_i)$ then we choose each $\boldsymbol{\sigma}_i$ using local information only. E.g., $\boldsymbol{\sigma}_i$ may be the sample variance computed on the basis of the nearest neighbors of $\mathbf{x}_i$. For the details of the nearest neighbor technique see [61]. Again the common scale h is chosen such that $p \in \mathscr{P}$.

4. Finally we can have $\Sigma_i = h \ \hat{\Sigma}_i$, where $\hat{\Sigma}_i$ is a full covariance matrix. Again each $\hat{\Sigma}_i$ should be chosen to be equal to the sample covariance derived from the neighborhood of the point $\mathbf{x}_i$ while the global scale h is again chosen so that the solution of the QPP gives a valid finite mixture pdf (82).

Case one is the only case for which we have an exact well-defined solution. Case two relies on a rigorously established asymptotic argument. The rest of the possibilities are well studied heuristics in kernel smoothing [56].

3.7 Solving the QPP

We comment on the solution of the QPP arising at each step of the GCE:

$$\begin{aligned} \min_{\mathbf{x}} \quad & \frac{1}{2}\mathbf{x}^T C\mathbf{x} - \mathbf{c}^T\mathbf{x} \\ \text{subject to:} \quad & \mathbf{x} \geqslant \mathbf{0}. \end{aligned}$$

This is a QPP subject to bound (box) constraints only. Note that this problem is one of the simplest problems in the class of QPPs. Since C is positive definite the KKT conditions are necessary and sufficient for a global solution $\mathbf{x}^*$. In this case the KKT conditions become:

$$\begin{aligned} C\mathbf{x}^* - \mathbf{c} - \boldsymbol{\pi}^* &= \mathbf{0} \\ \mathbf{x}^* &\geqslant \mathbf{0} \\ \boldsymbol{\pi}^* &\geqslant \mathbf{0} \\ \text{complementarity condition:} \quad \pi_i^* x_i^* &= 0 \quad \forall i, \end{aligned}$$

where $\boldsymbol{\pi}$ are the Lagrange multipliers associated with the constraint $\mathbf{x} \geqslant \mathbf{0}$. This is called the (monotone) *Linear Complementarity Problem* (LCP). Eliminating $\boldsymbol{\pi}$ gives:

$$\begin{aligned} \mathbf{x}^{*^T}(C\mathbf{x}^* - \mathbf{c}) &= 0 \\ \mathbf{x}^* &\geqslant \mathbf{0} \\ C\mathbf{x}^* &\geqslant \mathbf{c}. \end{aligned}$$

The LCP can be solved using the Wolfe-Danzig algorithm (see page 250 of [20]). Alternatively the system can be solved using Newton's method with a log-barrier penalty function taking care of the inequalities. This usually leads to a primal dual interior point method for the solution of the QPP. Interior point methods can solve the QPP in polynomial time. Numerical experience shows that the QPP does not cause any computational problems in terms of speed. The most computationally intensive part is calculating and storing the elements of the matrix C. For large problems, localized kernels, such as the uniform kernel, should be used to construct a sparse Hessian matrix C for the QPP.

Remark 9 (Log-barrier method) There is an alternative probabilistic view of the log barrier-interior point algorithm for solving the QPP. The solution of the problem:

$$\max_{\boldsymbol{\lambda}} \quad \sum_{k=1}^{n} \ln(\lambda_k) \tag{101}$$

$$\text{subject to:} \quad \frac{1}{2}\boldsymbol{\lambda}^T C\boldsymbol{\lambda} - \boldsymbol{\lambda}^T \mathbf{c} = r \tag{102}$$

approaches the solution of the QPP as the residual r is chosen smaller and smaller subject to existence of solutions (see [23]). We can interpret the above problem using the GCE postulate, namely, we are maximizing Burg's entropy of the distribution induced by the Lagrange multipliers $\boldsymbol{\lambda}$ subject to least squares fit to the observed data.

4 The Discrete GCE

In this section the GCE version for discrete stochastic optimization and machine learning is described. The general idea is still the same. Let $\mathscr{X}$ be a countable set of discrete states and let the probability mass function $q^* : \mathscr{X} \to [0,1]$, $\sum_{\mathbf{x}\in\mathscr{X}} q^*(\mathbf{x}) = 1$ be the target (possibly the Importance Sampling) pmf which solves a simulation or machine learning problem over the set $\mathscr{X}$. Then the prior pmf q is updated to p via the CE postulate with the ingredients:

1. Given the prior pmf q over the discrete set $\mathscr{X}$,
2. minimize the generalized Csiszár CE distance:

$$\min_{p \in \mathscr{P}} \mathcal{D}(p \to q),$$

where

(a) $\mathscr{P} \equiv \{p : \ p(\mathbf{x}) \geqslant 0, \ \sum_{\mathbf{x}} p(\mathbf{x}) = 1, \ \mathbf{x} \in \mathscr{X}\}$ is the set of all pmf's on $\mathscr{X}$,

(b) $\mathcal{D}(p \to q) = \sum_{\mathbf{x} \in \mathscr{X}} q(\mathbf{x}) \ \psi\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right)$,

(c) $\mathbf{x} \in \mathscr{X}$ is a column vector taking a countable number of discrete states,

3. subject to the *characterizing moment* constraints:

$$\mathbb{E}_p K_i(\mathbf{X}) = \sum_{\mathbf{x} \in \mathscr{X}} p(\mathbf{x}) \ K_i(\mathbf{x}) \geqslant \kappa_i^* \ , \quad i = 1, \ldots, n,$$

where

(a) κ_i^* is an estimate of $\mathbb{E}_{q^*} K_i(\mathbf{X})$,

(b) each $K_i : \mathscr{X} \to [0, 1]$ is a discrete unimodal kernel with the properties:

i. $\sum_{\mathbf{x} \in \mathscr{X}} K_i(\mathbf{x}) = 1$,

ii. $K_i(\mathbf{x}) \geqslant 0$ for all $\mathbf{x} \in \mathscr{X}$.

The dimension of the above optimization problem is equal to the size of the sample space and this space can be so large that a direct attack on the problem is impracticable. Again since ψ is strictly convex we use the theory of Lagrangian duality to reduce the dimension of the problem and find p. Let the **Primal Problem** be:

$$\min_{p} \quad \sum_{\mathbf{x} \in \mathscr{X}} q(\mathbf{x}) \ \psi\left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right) \tag{103}$$

$$\text{subject to:} \quad \sum_{\mathbf{x} \in \mathscr{X}} p(\mathbf{x}) \ K_i(\mathbf{x}) \ \geqslant \ \kappa_i^* \ , \quad i = 1, \ldots, n \tag{104}$$

$$\sum_{\mathbf{x} \in \mathscr{X}} p(\mathbf{x}) \ = \ 1 \tag{105}$$

$$p(\mathbf{x}) \ \geqslant \ 0, \quad \forall \mathbf{x} \in \mathscr{X}. \tag{106}$$

Unlike the continuous case, here we include the non-negativity constraint $p(\mathbf{x}) \geqslant 0$ in the definition of the primal problem. We now proceed to simplify

the notation. Since $\mathscr{X}$ is a countable discrete set we can put all the elements in $\mathscr{X}$ into one to one correspondence with the integers in $\{1, 2, \ldots, M\}$, where $M = |\mathscr{X}|$ could possibly be ∞. Note that for the GCE the order in which we label each of the states in $\mathscr{X}$ is irrelevant because $\mathcal{D}$ is permutationally symmetric (c.f. definition of generalized CE), as are all the constraints. Let $\mathbf{x}_m$ be the state corresponding to the integer m and $p_m = p(\mathbf{x}_m)$. The primal problem written in this new notation is:

$$\min_{p} \quad \sum_{m=1}^{M} q_m \, \psi\left(\frac{p_m}{q_m}\right) \tag{107}$$

$$\text{subject to:} \quad \sum_{m=1}^{M} p_m \, K_i(\mathbf{x}_m) \;\geqslant\; \kappa_i^* \,, \quad i = 1, \ldots, n \tag{108}$$

$$\sum_{m=1}^{M} p_m \;=\; 1 \tag{109}$$

$$p_m \;\geqslant\; 0, \quad m = 1, \ldots, M. \tag{110}$$

To derive the dual , define the Lagrangian:

$$\begin{aligned}
&\mathcal{L}(p; \boldsymbol{\lambda}, \boldsymbol{\eta}, \lambda_0) = \\
&= \sum_{m=1}^{M} \left(q_m \, \psi\left(\frac{p_m}{q_m}\right)\right) + \lambda_0 \left(1 - \sum_{m=1}^{M} p_m\right) + \sum_{i=1}^{n} \lambda_i \left(\kappa_i^* - \sum_{m=1}^{M} p_m \, K_i(\mathbf{x}_m)\right) - \sum_{m=1}^{M} \eta_m \, p_m \\
&= \lambda_0 + \sum_{i=1}^{n} \lambda_i \, \kappa_i^* + \sum_{m=1}^{M} \left(q_m \, \psi\left(\frac{p_m}{q_m}\right) - \lambda_0 \, p_m - \eta_m \, p_m - \sum_{i=1}^{n} \lambda_i \, p_m \, K_i(\mathbf{x}_m)\right) \\
&= \sum_{i=0}^{n} \lambda_i \, \kappa_i^* + \sum_{m=1}^{M} \left(q_m \, \psi\left(\frac{p_m}{q_m}\right) - \eta_m \, p_m - p_m \sum_{i=0}^{n} \lambda_i \, K_i(\mathbf{x}_m)\right) ,
\end{aligned}$$

where $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_n]^T$, $\boldsymbol{\eta} = [\eta_1, \ldots, \eta_M]^T$ and $\kappa_0^* = 1$, $K_0(\cdot) = 1$. Then the Wolfe **Dual Problem** is:

$$\max_{\boldsymbol{\lambda}, \boldsymbol{\eta}, \lambda_0} \quad \left\{ \inf_{p} \mathcal{L}(p; \boldsymbol{\lambda}, \boldsymbol{\eta}, \lambda_0) \right\} \tag{111}$$

$$\text{subject to:} \quad \boldsymbol{\lambda} \geqslant \mathbf{0}, \quad \boldsymbol{\eta} \geqslant \mathbf{0}. \tag{112}$$

If the constraints (104) were strict equalities instead of inequalities, the constraint $\boldsymbol{\lambda} \geqslant \mathbf{0}$ is omitted. We can find $\inf_p \mathcal{L}(p; \boldsymbol{\lambda}, \boldsymbol{\eta}, \lambda_0)$ from the first order necessary condition $\frac{\partial \mathcal{L}}{\partial p_m} = 0$, $m = 1, \ldots, M$. The unique solution is:

$$p_m = q_m \, \Psi' \left(\eta_m + \sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{x}_m) \right), \quad m = 1, \ldots, M.$$

Substituting this p into the Lagrangian and simplifying gives:

$$\begin{aligned}\mathcal{L}^*(\boldsymbol{\lambda}, \boldsymbol{\eta}, \lambda_0) &= \inf_p \mathcal{L}(p; \boldsymbol{\lambda}, \boldsymbol{\eta}, \lambda_0) \\ &= \sum_{i=0}^{n} \lambda_i \, \kappa_i^* + \sum_{m=1}^{M} q_m \, \psi\left(\Psi'\left(\nu_m + \sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{x}_m)\right)\right) \\ &\quad - \sum_{m=1}^{M} q_m \left(\eta_m + \sum_{i=0}^{n} \lambda_i \, K_i(\mathbf{x}_m)\right) \Psi'\left(\eta_m + \sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{x}_m)\right).\end{aligned}$$

Again use the equation $\Psi(x) = x\Psi'(x) - \psi\left(\Psi'(x)\right)$ to obtain:

$$\mathcal{L}^*(\boldsymbol{\lambda}, \boldsymbol{\eta}, \lambda_0) = \sum_{i=0}^{n} \lambda_i \, \kappa_i^* - \sum_{m=1}^{M} q_m \, \Psi\left(\eta_m + \sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{x}_m)\right).$$

Thus the simplest form of the **Dual Problem** is:

$$\max_{\boldsymbol{\lambda}, \boldsymbol{\eta}, \lambda_0} \quad \sum_{i=0}^{n} \lambda_i \, \kappa_i^* - \sum_{m=1}^{M} q_m \, \Psi\left(\eta_m + \sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{x}_m)\right) \tag{113}$$

$$\text{subject to:} \quad \boldsymbol{\lambda} \geqslant \mathbf{0}, \quad \boldsymbol{\eta} \geqslant \mathbf{0}. \tag{114}$$

Once we have a solution of the dual problem the solution of the primal is obtained via:

$$p_m = q_m \, \Psi'\left(\eta_m + \sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{x}_m)\right), \quad m = 1\ldots, M. \tag{115}$$

Similar to the continuous case we can argue that the simplest choice for ψ is $\psi'^{-1}(x) = x = \Psi'(x)$. Moreover we can again choose the kernels $\{K_i\}_{i=1}^n$ (e.g. by adjusting their respective scaling parameters) in such a way that the multipliers $\boldsymbol{\eta} = \mathbf{0}$, i.e., the constraint $p_m \geqslant 0, \ \forall m \Leftrightarrow p(\mathbf{x}) \geqslant 0, \ \mathbf{x} \in \mathscr{X}$ is inactive. Then the dual simplifies to:

$$\max_{\boldsymbol{\lambda}, \lambda_0} \quad -\frac{1}{2} + \sum_{i=0}^{n} \lambda_i \, \kappa_i^* - \frac{1}{2}\mathbb{E}_q \left(\sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{X})\right)^2 \tag{116}$$

$$\text{subject to:} \quad \boldsymbol{\lambda} \geqslant \mathbf{0}, \tag{117}$$

which is the same as

$$\max_{\boldsymbol{\lambda}, \lambda_0} \quad 2\sum_{i=0}^{n} \lambda_i \, \kappa_i^* - \sum_{i=0}^{n}\sum_{j=0}^{n} \lambda_i \, \lambda_j \, \mathbb{E}_q K_i(\mathbf{X}) K_j(\mathbf{X}) \tag{118}$$

$$\text{subject to:} \quad \boldsymbol{\lambda} \geqslant \mathbf{0}, \tag{119}$$

with primal solution:

$$p_m = p(\mathbf{x}_m) = q_m \sum_{k=0}^{n} \lambda_k \, K_k(\mathbf{x}_m), \quad m = 1, \ldots, M \, . \tag{120}$$

The dual can again be written in a convenient matrix notation:

$$\max_{\boldsymbol{\lambda}, \lambda_0} \quad 2 \, [\lambda_0, \boldsymbol{\lambda}^T] \begin{bmatrix} 1 \\ \boldsymbol{\kappa}^* \end{bmatrix} - [\lambda_0, \boldsymbol{\lambda}^T] \begin{pmatrix} 1 & \boldsymbol{\kappa}^T \\ \boldsymbol{\kappa} & B \end{pmatrix} \begin{bmatrix} \lambda_0 \\ \boldsymbol{\lambda} \end{bmatrix} \tag{121}$$

$$\text{subject to:} \quad \boldsymbol{\lambda} \geqslant \mathbf{0}, \tag{122}$$

where

$$B = \mathbb{E}_q \left[\mathbf{K}(\mathbf{X}) \, \mathbf{K}(\mathbf{X})^T \right].$$

Note that we have not eliminated λ_0 from the dual.

Choosing Discrete K

The simplest choice for a univariate discrete kernel (see [5] and [65]) on a finite discrete state space $\mathscr{D}$ is:

$$\mathcal{K}_i(x) = \mathcal{K}(x; x_i, \sigma_i) = \begin{cases} \sigma_i, & x = x_i \\ \frac{1-\sigma_i}{|\mathscr{D}|-1}, & x \neq x_i \end{cases}, \quad \frac{1}{|\mathscr{D}|} < \sigma_i \leqslant 1 \tag{123}$$

The restriction on the scale parameter σ_i guarantees that the kernel is unimodal and integrates to one. This kernel can be applied to both ordered and unordered categorical data. A multivariate kernel can easily be constructed as the product of univariate kernels:

$$K_i(\mathbf{x}) = K(\mathbf{x}; \mathbf{x}_i, \boldsymbol{\sigma}_i) = \prod_{l=1}^{d} \mathcal{K} \left(\mathbf{x}(l); \mathbf{x}_i(l), \boldsymbol{\sigma}_i(l) \right) \tag{124}$$

$$= \prod_{l=1}^{d} \boldsymbol{\sigma}_i(l)^{I\{\mathbf{x}(l)=\mathbf{x}_i(l)\}} (1 - \boldsymbol{\sigma}_i(l))^{1-I\{\mathbf{x}(l)=\mathbf{x}_i(l)\}}, \tag{125}$$

where $\boldsymbol{\sigma}_i = [\sigma_i(1), \ldots, \sigma_i(d)]^T$ is a vector of bandwidth parameters associated with each dimension of $\mathbf{x}$. If for simplicity we assume that $\boldsymbol{\sigma}_i = \sigma \, [1, \ldots, 1]^T$, then:

$$K_i(\mathbf{x}) = \prod_{l=1}^{d} \sigma^{I\{\mathbf{x}(l)=\mathbf{x}_i(l)\}} \left(\frac{1-\sigma}{|\mathscr{D}|-1} \right)^{1-I\{\mathbf{x}(l)=\mathbf{x}_i(l)\}} \tag{126}$$

$$= \sigma^{\sum_l^d I\{\mathbf{x}(l)=\mathbf{x}_i(l)\}} \left(\frac{1-\sigma}{|\mathscr{D}|-1} \right)^{d-\sum_{l=1}^d I\{\mathbf{x}(l)=\mathbf{x}_i(l)\}} \tag{127}$$

$$= \sigma^{\mathrm{d}(\mathbf{x};\mathbf{x}_i)} \left(\frac{1-\sigma}{|\mathscr{D}|-1} \right)^{d-\mathrm{d}(\mathbf{x};\mathbf{x}_i)}, \tag{128}$$

where $\mathrm{d}(\mathbf{x};\mathbf{y}) = \sum_{l=1}^{d} I\{\mathbf{x}(l) = \mathbf{y}(l)\}$.

Example 14 (Kernel for Binary Data) Suppose that $\mathbf{x}$ is a binary vector, i.e., $|\mathscr{D}| = 2$, then the binary kernel with a single bandwidth parameter is:

$$K_i(\mathbf{x}) = \sigma^{\mathrm{d}(\mathbf{x};\mathbf{x}_i)}(1-\sigma)^{d-\mathrm{d}(\mathbf{x};\mathbf{x}_i)}, \quad \frac{1}{2} < \sigma \leqslant 1.$$

Then for a uniform prior q:

$$\begin{aligned}
\sum_{\mathbf{x}\in\mathscr{X}} K_i(\mathbf{x})\, K_j(\mathbf{x}) &= \sum_{\mathbf{x}\in\mathscr{X}} \prod_{l=1}^{d} \sigma^{I\{\mathbf{x}(l)=\mathbf{x}_i(l)\}+I\{\mathbf{x}(l)=\mathbf{x}_j(l)\}}(1-\sigma)^{2-I\{\mathbf{x}(l)=\mathbf{x}_i(l)\}-I\{\mathbf{x}(l)=\mathbf{x}_j(l)\}} \\
&= \prod_{l=1}^{d} \sum_{\mathbf{x}(l)} \sigma^{I\{\mathbf{x}(l)=\mathbf{x}_i(l)\}+I\{\mathbf{x}(l)=\mathbf{x}_j(l)\}}(1-\sigma)^{2-I\{\mathbf{x}(l)=\mathbf{x}_i(l)\}-I\{\mathbf{x}(l)=\mathbf{x}_j(l)\}} \\
&= \prod_{l=1}^{d} \Big(I\{\mathbf{x}_i(l) = \mathbf{x}_j(l)\} \left[\sigma^2 + (1-\sigma)^2\right] + I\{\mathbf{x}_i(l) \neq \mathbf{x}_j(l)\}\, 2\sigma(1-\sigma) \Big) \\
&= \left[\sigma^2 + (1-\sigma)^2\right]^{\mathrm{d}(\mathbf{x}_i;\mathbf{x}_j)} [2\sigma(1-\sigma)]^{d-\mathrm{d}(\mathbf{x}_i;\mathbf{x}_j)} \\
&= \varsigma^{\mathrm{d}(\mathbf{x}_i;\mathbf{x}_j)}(1-\varsigma)^{d-\mathrm{d}(\mathbf{x}_i;\mathbf{x}_j)}, \quad \varsigma = \sigma^2 + (1-\sigma)^2.
\end{aligned}$$

We can thus compute $B_{ij} = \mathbb{E}_q K_i(\mathbf{X}) K_j(\mathbf{X})$ without too much trouble.

Kernels living on an infinite countable state space can be envisioned (see [5]). We can also construct mixed kernels which combine discrete and continuous spaces in a product kernel form. They could possibly be used in the simulation of non-Markovian stochastic jump processes.

Sampling from p, the estimation of $\boldsymbol{\kappa}^*$ and the solution of the associated QPP is analogous to the continuous case.

As a consequence of using Pearson's χ^2 CE distance $\mathscr{D}_2$ in the GCE method we have the following useful result from [33].

Theorem 17 Suppose at a certain iteration we have n random observations from the prior $q(\mathbf{x})$ (which approximates the target $q^*(\mathbf{x})$). Let $E_m = n \times q_m$ denote the expected number of observations under the prior and $O_m = n \times p_m$ the (observed) frequency under the estimated GCE pmf, then $2n \times \mathscr{D}_2(p \to q)$ has approximately a χ^2 distribution with $M - n - 1$ degrees of freedom:

$$\begin{aligned}
2n\, \mathscr{D}_2(p \to q) &= 2n \frac{1}{2} \sum_{m=1}^{M} \frac{(p_m - q_m)^2}{q_m} && (129) \\
&= \sum_{m=1}^{M} \frac{(np_m - nq_m)^2}{nq_m} && (130) \\
&= \sum_{m=1}^{M} \frac{(O_m - E_m)^2}{E_m} \overset{approx.}{\sim} \chi^2_{M-n-1}. && (131)
\end{aligned}$$

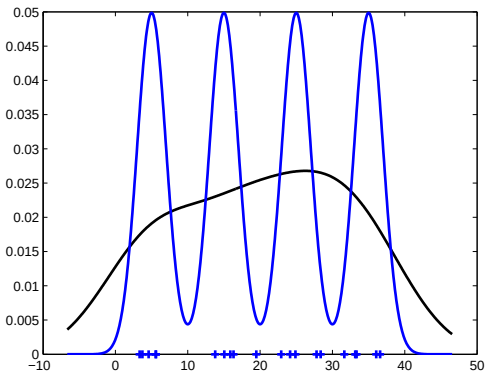
With the inclusion of the normalizing constraint we have $n + 1$ constraints in the primal which reduce the degrees of freedom from M to $M - n - 1$.

The result is only asymptotic but can still be used to conduct a χ^2 goodness-of-fit test. The test can establish whether any difference between the prior $q(\mathbf{x})$ and the updated $p(\mathbf{x})$ is statistically significant. If the difference is not significant then we terminate any iterative updating of $p(\mathbf{x})$ and treat $p(\mathbf{x})$ as a reasonable approximation of $q^*(\mathbf{x})$.

5 Application to Data Modeling

In this section we apply the GCE method to the problem of probability density estimation. Recall the main problem of statistical learning: Given a finite number of empirical observations $\mathcal{X}_n = (\mathbf{X}_1, \ldots, \mathbf{X}_n)$, find an optimal in some sense model for $\mathcal{X}_n$ using as few assumption as possible. Even more abstractly the problem of learning is to estimate a (density) function from a finite number of observations/specifications. A function is, for all practical purposes or most of the time, an infinite dimensional object. The specifications (empirical data say), however, are finite in number. Intuitively we know that there is no way of obtaining an infinite dimensional object from a finite number of specifications unless we introduce extra information and assumptions in the model. Learning is thus an ill-posed problem — a problem which does not have a unique and stable solution. Ill-posed problems are usually solved using Regularization Theory (see [66]). This theory imposes in a systematic way the fewest/weakest assumptions necessary for a unique stable and well-behaved solution of the problem to exist.

Example 15 (Statistical modeling — an ill-posed problem) To get an idea of why data modeling is an ill-posed problem suppose we are given one dimensional continuous data on $\mathbb{R}$ (see graph below). What is the best possible probabilistic model for the data? The black probability density function with a single bump or the blue multi-modal pdf?



In both cases the data points are represented by the blue plus signs at the bottom of the graphs. There are reasons to prefer the simpler and sparser model. In our case we might prefer the simpler black versus the more complicated blue multi-modal density. We may argue that the data is not numerous enough to justify multiple modes. As a matter of fact the data was generated synthetically (using MATLAB's random number generator) from the blue mixture pdf, yet the black curve which represents the current state of the art in density estimation is not even multi-modal. This is partly what makes the problem ill-posed. We may have reasons to prefer the black curve but the blue curve is also a reasonable model for the data. The data simply does not provide enough information to give a unique or well defined solution to the density estimation problem.

Now that we have stated the gist of problem we briefly review the approaches taken so far toward resolving this problem.

5.1 Classical Approach to Statistical Learning

The approaches to statistical modeling have so far been quite unsatisfactory. In the example above we may specify a function subjectively up to a small number of parameters (say Gaussian with mean μ and variance σ^2— $\mathsf{N}(\mu, \sigma^2)$) and estimate these parameters. The focus of classical statistics is on estimating/finding the few model parameters (say (μ, σ)) in an optimal way. This problem has been largely solved by Sir Ronald Fisher in the beginning of the twentieth century. He gave the likelihood principle as the asymptotically efficient estimation method. In the classical paradigm one has to specify the probability density function **subjectively** and then proceed to estimate the parameters in a rigorous way! This approach is usually referred to as the parametric approach to statistics — it focuses on optimal parameter estimation. The major drawback of this approach is that an incorrectly specified parametric function does not necessarily converge to the unknown density function q^* as the sample size grows to infinity. Moreover it is hard to verify the validity of the parametric model assumptions[9] and small perturbations of the parametric assumptions render the Maximum Likelihood Estimators (e.g., sample mean and variance are Maximum Likelihood Estimators of μ and σ^2) asymptotically inefficient. So unless we have prior knowledge about the correctness of the assumptions, the classical approach is bound to fail. The subjectivism in choosing the functional form of the probability density and the rigor with which one estimates the parameters of the density has prompted the famous statistician Tukey to articulate his concern by saying "It is better to be approximately right than exactly wrong!". In fact the subjectivism of the classical approach has been the reason mathematicians from other fields dismiss Statistics as nonsense.

[9] [56] argues that with large samples, goodness-of-fit test almost always reject quite reasonable models.

5.2 The Non-Parametric Approach

The focus has recently shifted on directly estimating the entire probability density function, not just a few parameters of a subjectively specified function (see [56]). This idea was first advocated by a contemporary of Fisher — Karl Pearson — and is usually referred to as non-parametric statistics to stress the fact that the focus is no longer on optimal parameter estimation[10]. Pearson's idea did not gain popularity because the non-parametric approach to the problem of learning is computer intensive relative to the classical approach. So far researchers have favored the classical approach due to its computational simplicity, but the recent explosion of computing power has made Pearson's ideas workable and indeed very competitive to the classical approach. The non-parametric approach takes on a more direct attack on the learning problem. More specifically it tries to approximately solve an infinite dimensional functional problem to find the functional relationship (e.g., a probability density function) that best describes the pattern in the empirical data. The resulting density estimate converges to the unknown density q^* as the number of empirical observations grow to infinity. This is what makes the non-parametric approach consistent. The price to pay for removing the subjective element in nonparametric Statistics is enormous computational complexity compared to the simplicity of parametric Statistics. Currently the most popular non-parametric approach to density estimation is the kernel approach (for a general introduction see [56], [68], [62]) with its many different flavors (see, e.g., [14], [46], [44], [55], [43], [64], [3]). We review this technique before presenting our GCE solution.

5.3 The Kernel Approach to Learning

Suppose we are given $d-$dimensional data $\mathcal{X}_n \equiv \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ on $\mathbb{R}^d$ and wish to visualize any patterns present in it, compress it or draw inferences based on nonparametric statistical analysis. We wish to model the data probabilistically. Assume that the data is the random outcome of an unknown continuous probability density function $q^*(\mathbf{x})$, i.e.:

$$\mathbf{X}_1, \ldots, \mathbf{X}_n \overset{i.i.d}{\sim} q^*(\mathbf{x}).$$

Thus the problem is to find/estimate q^* using the empirical data and as few assumptions as possible. We can summarize all the information present in the data via the empirical pdf

$$\Delta(\mathbf{t}) = \frac{1}{n} \sum_{i=1}^{n} \delta(\mathbf{t} - \mathbf{X}_i).$$

[10]Fisher and Pearson—the two major proponents of the classical and non-parametric approaches respectively—were bitter adverseries in a way similar to Newton and Leibniz.

Since $\Delta(\mathbf{t})$ is not continuous, it is useless as an estimate of the continuous and possibly differentiable q^*. To "smooth" the atomic and discontinuous density $\Delta(\mathbf{t})$ we can borrow the convolution method for smoothing "rough" and "noisy" signals. The idea is to convolve $\Delta(\mathbf{t})$ with a suitable continuous function $K_h : \mathbb{R}^d \to \mathbb{R}^+$ depending on a parameter h which controls the amount of "smoothing" applied to the spiky "signal" $\Delta(\mathbf{t})$. This procedure leads to the "smooth" estimate of q^*:

$$\begin{aligned} f(\mathbf{x}\,|\,h, \mathcal{X}_n) &= K_h(\mathbf{t}) * \Delta(\mathbf{t})\,[\mathbf{x}] \\ &= \int_{\mathbb{R}^d} K_h(\mathbf{x}-\mathbf{t})\,\frac{1}{n}\sum_{i=1}^{n}\delta(\mathbf{t}-\mathbf{X}_i)\,d\mathbf{t} = \frac{1}{n}\sum_{i=1}^{n} K_h(\mathbf{x}-\mathbf{X}_i). \end{aligned}$$

We can now use the smoothing parameter h to minimize a suitable measure of distance between our proposed model $f(\mathbf{x}\,|\,h, \mathcal{X}_n)$ and the desired target q^*. Thus the idea of convolution from signal processing motivates the so called kernel method. The method, similar to the Rayleigh Ritz method, assumes that the true, but unknown, underlying density function q^* can be approximated well by a probability density function[11] of the form:

$$f(x\,|\,h, \mathcal{X}_n) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right), \tag{132}$$

where:

1. $h \in \mathbb{R}^+\backslash\{0\}$ is a bandwidth parameter which controls the "smoothness" or "resolution" of f in a way similar to the convolution operation in signal processing.

2. $K : \mathbb{R} \to \mathbb{R}^+$, $\int_{\mathbb{R}} K(x)\,dx = 1$, $K(-x) = K(x)$, i.e., K is a symmetric unimodal kernel. For our purposes we choose to use the Gaussian kernel $K(x) = \phi(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}x^2}$.

3. $X_1, \ldots, X_n \overset{i.i.d}{\sim} q^*(x)$, i.e., we assume the data can be modeled as the outcome of a random experiment with density function q^*.

The idea behind the kernel method is that just like a Taylor series (a set of polynomial functions $\{(x-a)^i\}_{i=0}^n$) or a Fourier series (a set of orthogonal functions $\{\sin(nx), \cos(nx)\}_{i=0}^n$) can represent many functions arbitrarily well, so can the kernel set $\left\{K\left(\frac{x-X_i}{h}\right)\right\}_{i=1}^n$ represent a quite general density q^* very well. Note that this assumption is much weaker than the assumptions of the parametric approach.

Everything in (132) is fixed except the bandwidth h. This is the only parameter over which we have control. We now need to tune h so that our approximation of q^* is as good as possible.

[11] For simplicity assume that $d = 1$, i.e., the data is one-dimensional.

5.4 Measuring the performance/error

Once we have defined the class of functions within which we search for the (best in some sense) solution we now have to choose a measure of performance. In other words we have to choose a measure of distance between the proposed model (132) and the observed empirical data. Classical statistics gives the Mean Squared Error (MSE) as a measure of the performance of various estimators. We choose to work with the MSE criterion due to its computational tractability:

$$\mathsf{MSE}\{f\}(x\,|\,h) = \mathbb{E}_{q^*}\left[f(x\,|\,h,\mathcal{X}_n) - q^*(x)\right]^2. \tag{133}$$

We can write (133) as:

$$\mathsf{MSE}\{f\}(x\,|\,h) = \underbrace{\left[\mathbb{E}_{q^*} f(x\,|\,h,\mathcal{X}_n) - q^*(x)\right]^2}_{\mathsf{Bias}^2(x\,|\,h)} + \underbrace{\mathbb{E}_{q^*}\left[f(x\,|\,h,\mathcal{X}_n)\right]^2 - \left[\mathbb{E}_{q^*} f(x\,|\,h,\mathcal{X}_n)\right]^2}_{\mathrm{Var}(x\,|\,h)}. \tag{134}$$

Each of the components of MSE above can be simplified using the i.i.d assumption[12]:

$$\mathsf{Bias}(x\,|\,h) = \frac{1}{h}\int K\left(\frac{x-z}{h}\right) q^*(z)\,dz \;\; - q^*(x) = \int K(z)\; q^*(x - hz)\,dz - q^*(x), \tag{135}$$

$$\mathsf{Var}(x\,|\,h) = \frac{1}{nh}\int K^2(z)\; q^*(x - hz)\,dz - \frac{1}{n}\left[\int K(z)\; q^*(x - hz)\,dz\right]^2. \tag{136}$$

Therefore, dropping f from the MSE notation:

$$\begin{aligned}\mathsf{MSE}(x\,|\,h) = \frac{1}{nh}\int K^2(z)\; q^*(x - hz)\,dz + (1 - n^{-1})\left[\int K(z)\; q^*(x - hz)\,dz\right]^2 \\ -2\; q^*(x)\int K(z)\; q^*(x - hz)\,dz + [q^*(x)]^2.\end{aligned}$$

We can now minimize the MSE for each given value of x by tweaking the parameter h, i.e.:

$$\min_{h>0}\; \mathsf{MSE}(h\,|\,x).$$

The h which minimizes the MSE for each x, say $h^*(x)$, is a function of x itself. Rather than estimating the unknown q^* at each point x, we wish to have a single value for h, say h^*, which globally minimizes the discrepancy between f and q^*. One convenient measure of 'goodness of fit' over the entire real line is the Mean Integrated Squared Error:

$$\mathsf{MISE}\{f\}(h) = \mathbb{E}_{q^*}\left[\int \left[f(x\,|\,h,\mathcal{X}_n) - q^*(x)\right]^2 dx\right] = \int \mathbb{E}_{q^*}\left[f(x\,|\,h,\mathcal{X}_n) - q^*(x)\right]^2 dx. \tag{137}$$

[12] integration taken over entire real line

MISE is simply the accumulated pointwise MSE error across the real line:

$$\mathsf{MISE}\{f\}(h) = \int \mathsf{MSE}(x \,|\, h)\, dx.$$

Whence (again omitting f from MISE):

$$\begin{aligned}\mathsf{MISE}(h) \;&=\; \frac{1}{nh}\int K^2(z)\, dz + (1+n^{-1})\int\left[\int K(z)\, q^*(x-hz)\, dz\right]^2 dx \\ &-\; 2\int q^*(x)\int K(z)\, q^*(x-hz)\, dz\, dx + \int [q^*(x)]^2\, dx.\end{aligned}$$

We now have two measures of discrepancy between f and q^* - one global (MISE) and one pointwise local (MSE). Each of these measures will give a different optimal bandwidth, namely: $h^*(x)$ and h^*. The first bandwidth is a function of x and will be different at each point of estimation, the second is constant across the real line. Notice that $\int_{\mathbb{R}} f(x \,|\, h^*(x), \mathcal{X}_n)\, dx$ does not in general equal one while $\int_{\mathbb{R}} f(x \,|\, h^*, \mathcal{X}_n)\, dx = 1$ always. Naturally our density estimate f has to be a proper pdf and hence integrate to one. Thus MISE is the criterion of choice for our subsequent discussion. Finding an optimal h thus reduces to the following program:

$$h^* = \min_{h>0}\ \mathsf{MISE}(h). \tag{138}$$

Finding a unique and explicit solution to (138) is impossible due to the complicated nature of the integrals appearing in MISE and the fact that q^* is unknown — only a few random realizations from it are given. Using large sample (asymptotic) theory (see [56]) we can, however, obtain a unique explicit answer which approximates the solution to (138).

5.5 Asymptotic Expansion of MISE

We will explore the behavior of MISE as the sample size grows larger and larger, i.e., as $n \to \infty$. In the large sample analysis, we use the following crucial assumptions:

1. The bandwidth h depends on the size of the sample $\mathcal{X}_n$ in such a way that:
$$\lim_{n\to\infty} h_n = 0.$$
2. The rate at which the optimal bandwidth goes to zero is smaller than $O(n^{-1})$, i.e.:
$$\lim_{n\to\infty} n \times h_n = \infty.$$
3. $\frac{d^2}{dx^2}\,(q^*(x))$ is continuous, square integrable function.

These conditions are borrowed from the "General Kernel Density Estimator" theorem in the preliminary section. They ensure that f is a consistent non-parametric density estimator. Using assumptions 1. and 3., the symmetric property $\int z\, K(z)\, dz = 0$ and Taylor's expansion of $q^*(x - zh)$ about x, equation (135) becomes:

$$\mathsf{Bias}(x\,|\,h) = \frac{h^2}{2}\, q^{*''}(x) \int z^2 K(z)\, dz + o\left(h^3\right), \quad n \to \infty. \tag{139}$$

The bias depends on the curvature of q^* and regions with high curvature, i.e., large $q^{*''}(x)$, are difficult to estimate. Note that asymptotically the pointwise bias does not depend on n and increasing n alone will not reduce the bias unless $h_n \to 0^+$ as $n \to \infty$. Equation (136) can be similarly expanded:

$$\mathrm{Var}(x\,|\,h) = \frac{q^*(x)}{nh} \int K^2(z)\, dz + o\left(\frac{1}{nh}\right), \quad n \to \infty. \tag{140}$$

Note that $\mathrm{Var}(x\,|\,h) \to 0$ as $n \to \infty$ under assumption 2. Thus the asymptotic expansions of MSE and MISE are :

$$\mathsf{MSE}(x\,|\,h) = \underbrace{\frac{q^*(x)}{nh} \int K^2(z)\, dz + \frac{h^4}{4}\left[q^{*''}(x) \int z^2 K(z)\, dz\right]^2}_{\mathsf{AMSE}(x\,|\,h)} + o\left(h^4 + \frac{1}{nh}\right), \tag{141}$$

$$\begin{aligned} \mathsf{AMISE}(h) &= \int \mathsf{AMSE}(x\,|\,h)\, dx & (142)\\ &= \frac{\int K^2(z)\, dz}{nh} + \frac{h^4\left[\int z^2\, K(z)\, dz\right]^2}{4} \int \left[q^{*''}(z)\right]^2\, dz, & (143) \end{aligned}$$

where AMSE stands for Asymptotic Mean Squared Error and AMISE stands for Asymptotic Mean Integrated Squared Error. AMISE gives the first order asymptotic behavior of MISE as the sample size grows to infinity. What makes AMISE attractive is that the program:

$$\min_{h>0} \mathsf{AMISE}(h)$$

can be solved explicitly to give the optimal asymptotic bandwidth:

$$h_{\mathsf{AMISE}} = \left[\frac{\int K^2(z)\, dz}{n\left[\int z^2\, K(z)\, dz\right]^2 \int \left[q^{*''}(z)\right]^2 dz}\right]^{1/5}. \tag{144}$$

Apart from its dependence on the known kernel K and n, h^5_{AMISE} is inversely proportional to $\int [q^{*''}(x)]^2 dx$. The functional $\int [q^{*''}(x)]^2 dx$ measures the total curvature of q^*. Thus for densities with little curvature a large bandwidth will be required. Alternatively when $\int [q^{*''}(x)]^2 dx$ is large, little smoothing will be optimal. The value h_{AMISE} gives the minimum of AMISE:

$$\begin{aligned} \mathsf{AMISE}(h_{\mathsf{AMISE}}) &= \min_{h>0} \mathsf{AMISE}(h) \\ &= \frac{5}{4} n^{-4/5} \left(\left[\int K^2(z)\, dz \right]^4 \left[\int z^2\, K(z)\, dz \right]^2 \int \left[q^{*''}(z) \right]^2 dz \right)^{1/5}. \end{aligned}$$

Notice that $\mathsf{AMISE}(h_{\mathsf{AMISE}}) \to 0$ as $n \to \infty$ at the rate of $n^{-4/5}$. Thus our estimator indeed converges to the target. It seems like we have solved the problem of density estimation, at least in the large sample case. Unfortunately the optimal asymptotic bandwidth h_{AMISE} still depends on the unknown density q^* through the functional $\int \left[q^{*''}(z) \right]^2 dz$. Thus h_{AMISE}, which is an approximation to h^*, needs to be estimated. Almost all of the current hi-tech bandwidth selection methods (see [68]) use a variation of the so called *plug-in* method in which the functional $\int [q^{*''}(z)]^2 dz$ is estimated using a rough pilot density estimate and then the resulting estimate is substituted into (144) to obtain an approximation to h_{AMISE}. Note that this approach is a long way from our initial target, namely solving:

$$\min_{h>0} \mathsf{MISE}(h).$$

The trouble is that the plug-in method gives a bandwidth which approximately minimizes an asymptotic approximation of the MISE! This is not desirable but we have very few options and in practice these approximations work well for reasonably large n.

5.6 The Sheather-Jones plug-in bandwidth estimate

Arguably the best data-driven bandwidth selection method is the plug-in Sheather Jones bandwidth estimator (see [45] and [60]). Here we will give the gist of the method without going into the details.

The functional $\int [q^{*''}(z)]^2\, dz = \int q^*(z)\, q^{*(4)}(z)\, dz = \mathbb{E}_{q^*}[q^{*(4)}(X)]$ using straightforward integration by parts and assuming that q^* is four times differentiable. The function $q^{*(4)}(x)$ can be estimated by the forth derivative of the kernel estimator:

$$f^{(4)}(x \,|\, \alpha, \mathcal{X}_n) = \frac{1}{\alpha^5 n} \sum_{i=1}^{n} K^{(4)}\left(\frac{x - X_i}{\alpha} \right).$$

Notice that the bandwidth used to estimate $q^{*(4)}(x)$ is α, not h. We will come to this issue later. The expectation $E_{q^*}[q^{*(4)}(X)]$ is approximated by an empirical

average [13]:

$$\int \left[q^{*''}(z)\right]^2 \, dz = \mathbb{E}_{q^*}[q^{*(4)}(X)] \tag{145}$$

$$\approx \frac{1}{n-1} \sum_{i=1}^{n} f^{(4)}(X_i \,|\, \alpha, \mathcal{X}_n) \tag{146}$$

$$= \frac{1}{\alpha^5 \; n \; (n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} K^{(4)}\left(\frac{X_i - X_j}{\alpha}\right) = S[\alpha]. \tag{147}$$

Intriguingly the optimal value of α used to estimate the forth derivative of q^* is different from the optimal value of h used to estimate q itself. In an intricate asymptotic argument Sheather and Jones [60] establish an optimal asymptotic relationship between α and h. They find that if h_{AMISE} is the bandwidth that achieves the best asymptotic estimate for $q^*(x)$, then $\alpha_{\mathsf{AMISE}} = c \; (h_{\mathsf{AMISE}})^{5/7}$ will be the bandwidth that best estimates $q^{*(4)}(x)$ asymptotically. c is generally an unknown constant. Sheather and Jones suggest a heuristic choice for c (for the details see [60]) based on a rough pilot estimate of q^*. Finally the optimal **Sheather-Jones kernel estimator** is obtained from:

1. Solve numerically the equation (an approximation to (144)):

$$h - \left[\frac{\int K^2(z) \, dz}{n \left[\int z^2 \; K(z) \, dz\right]^2 S[\alpha(h)]}\right]^{1/5} = 0, \tag{148}$$

   where $S[\alpha] = \frac{1}{\alpha^5 \; n \; (n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} K^{(4)}\left(\frac{X_i - X_j}{\alpha}\right)$ and $\alpha = c \times h^{5/7}$, the constant c being a judiciously chosen number. The solution of the equation gives the optimal Sheather-Jones bandwidth h_{SJ}.

2. Present the equally weighted (Gaussian in our case) mixture pdf:

$$f(x \,|\, h_{\mathsf{SJ}}, \mathcal{X}_n) = \frac{1}{n \, h_{\mathsf{SJ}}} \sum_{i=1}^{n} \phi\left(\frac{x - X_i}{h_{\mathsf{SJ}}}\right), \tag{149}$$

   where $K(x) = \phi(x) = \frac{1}{\sqrt{2\pi}} \, \mathrm{e}^{-x^2/2}$, as the kernel density model for the data $\mathcal{X}_n$.

Numerical experiments demonstrating the performance of the Sheather-Jones bandwidth selection method are presented in the final section. Now that we have introduced the current mainstream method for density estimation we present the alternative GCE approach to the problem of density estimation.

[13] except that we divide by $(n-1)$ and not n;

5.7 Density Estimation via GCE

For clarity we now restate the crux of the GCE method in the context of one-dimensional density estimation ($d = 1$). Again assume that all we have is the empirical data $\mathcal{X}_n = \{X_1, \ldots, X_n\}$. Then apply the GCE postulate with the following elements:

1. Given the uniform/uninformative prior $q \propto 1$ on $\mathbb{R}$,
2. solve the functional optimization program:

$$\min_{p \in \mathscr{P}} \mathcal{D}_2(p \to q) \equiv \min_{p \in \mathscr{P}} \int_{\mathbb{R}} p^2(x)\, dx, \tag{150}$$

3. subject to the constraint set $\mathscr{C}$:

$$\int_{\mathbb{R}} p(x)\, K_i(x)\, dx = \mathbb{E}_p[K_i(X)] \geqslant \kappa_i^* = \frac{1}{n-1} \sum_{j \neq i} K_i(X_j), \quad i = 1, \ldots, n. \tag{151}$$

Again $\mathscr{P} = \left\{p : \int_{\mathbb{R}} p(x)\, dx = 1,\ p(x) \geqslant 0,\ x \in \mathbb{R}\right\}$ denotes the set of all probability density functions on $\mathbb{R}$ and, just like in the kernel method, we choose a Gaussian kernel $K_i(x) = K(x; X_i, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-X_i)^2}{2\sigma^2}\right) = \frac{1}{\sigma}\phi\left(\frac{x-X_i}{\sigma}\right)$. We can interpret the program $\min_{p \in \mathscr{P}} \mathcal{D}_2$ as minimization of the complexity of the proposed probabilistic model p and the imposition of the constraint set $\mathscr{C}$ as a means of ensuring that the model is consistent with the empirical data. The above problem is equivalent to the dual formulation:

1. Solve the program:

$$(\sigma^*, \boldsymbol{\lambda}^*) = \left\{(\sigma, \boldsymbol{\lambda}) :\ \mathbf{1}^T \boldsymbol{\lambda}(\sigma) = 1,\ \boldsymbol{\lambda}(\sigma) = \operatorname*{argmin}_{\boldsymbol{\lambda} \geqslant \mathbf{0}} \left(\frac{1}{2}\boldsymbol{\lambda}^T C(\sigma)\boldsymbol{\lambda} - \boldsymbol{\lambda}^T \boldsymbol{\kappa}^*(\sigma)\right)\right\}, \tag{152}$$

   where the matrix $C_{n \times n}$ has entries $C_{ij} = \int_{\mathbb{R}} K_i(x; X_i, \sigma)\, K_j(x; X_j, \sigma)\, dx = \frac{1}{\sqrt{2}\sigma}\phi\left(\frac{X_i - X_j}{\sqrt{2}\sigma}\right) = \frac{1}{\sqrt{2\pi}(\sqrt{2}\sigma)} \exp\left(-\frac{(X_i - X_j)^2}{4\sigma^2}\right)$.

2. Present the Gaussian mixture pdf

$$p(x) = \sum_{j=1}^{n} \lambda_j^*\, K(x; X_j, \sigma^*) \tag{153}$$

   as the optimal GCE density which models the data $\mathcal{X}_n$.

6 Numerical Experiments

In this section we present some numerical experiments demonstrating the performance of the Sheather-Jones and GCE probability density estimators.

Matlab Implemenation

Some issues concerning the implementation of the Sheather-Jones method and the GCE method are :

1. The Matlab routine used in our simulation experiments implementing the Sheather-Jones bandwidth method was downloaded from Professor Steve Marron's website:

   http://www.stat.unc.edu/faculty/marron/marron_software.html

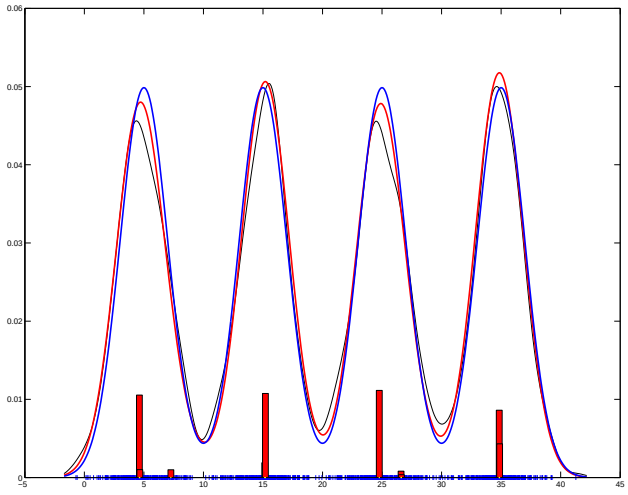
2. The compiled Matlab routine "mosekopt" is used to solve the QPP in the GCE optimization. "mosekopt" was downloaded from this webpage:

   http://www.mosek.com/trials.html# students

3. To solve the program (152) we use the Matlab build-in root finding function "fzero.m". Each iteration of "fzero.m" requires the solution of a QPP and hence calls "mosekopt".

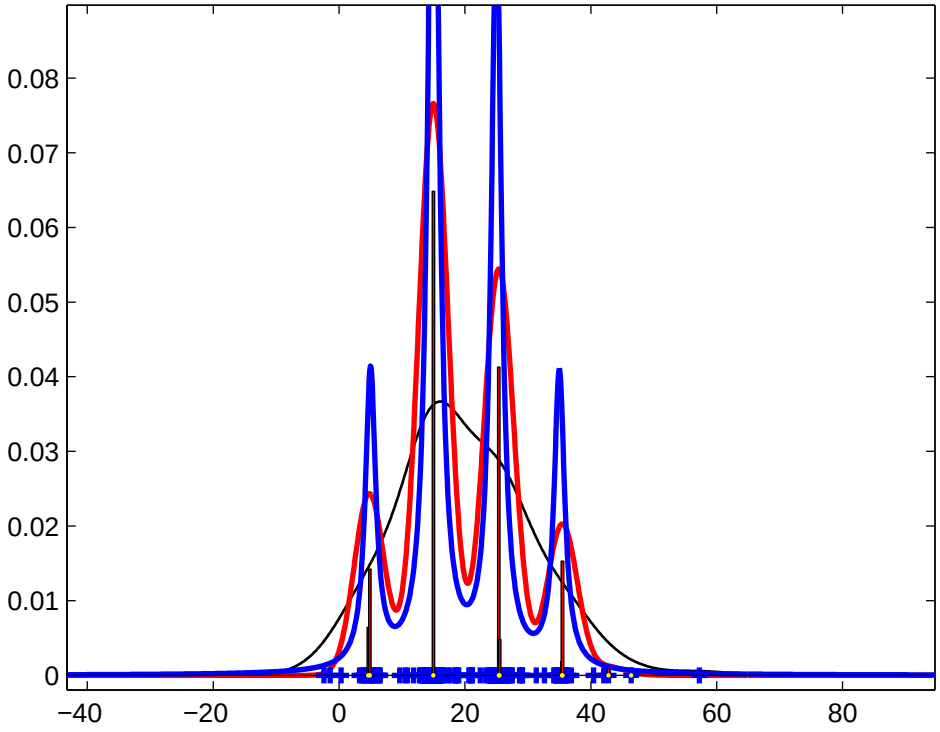
We now present some density estimation examples with synthetically generated data. The data was generated using Matlab's random number generator.

Example 16 (Gaussian mixture) We consider the following model. 1020 points were generated from an equally weighted mixture of Gaussians with a common scale parameter. The mixture is given by the blue curve on the graph below and the points are represented as crosses on the real line.

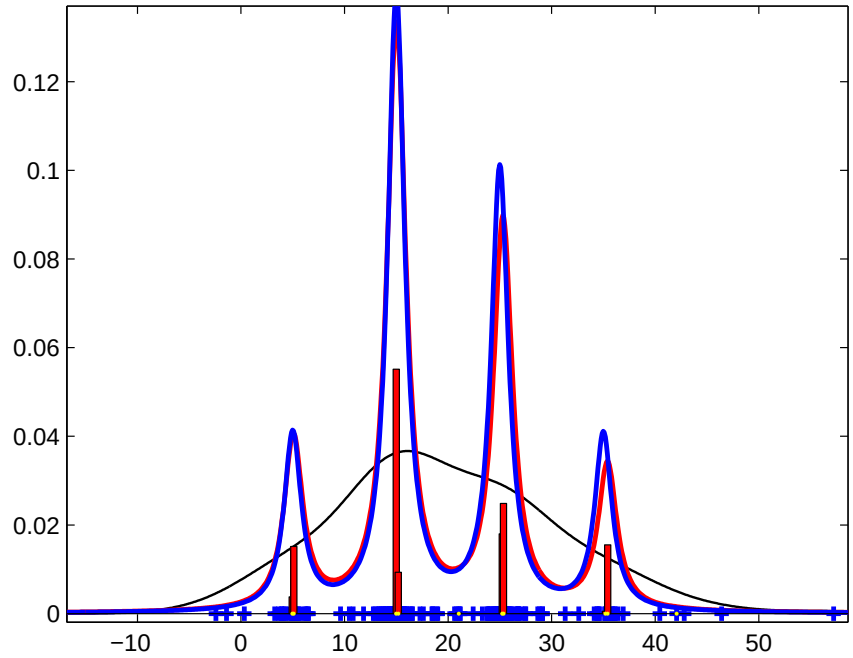


The black curve is the Sheather-Jones estimator (149). The red curve represents the GCE estimator (153). The red bars represent the relative values of the Lagrange multipliers λ (i.e., mixture weights of (153)) associated with each point. It is interesting to note that out of the 1020 points only 10 points have non-zero Lagrange multipliers. Thus the GCE model for the 1020 points is a Gaussian mixture with 10 components only. In contrast, the Sheather-Jones estimator is an equally weighted mixture with 1020 components. The sparsity of the GCE estimator makes it computationally easier to evaluate at each point and visualize. Apart from this there is not much difference in the performance of the two estimators.

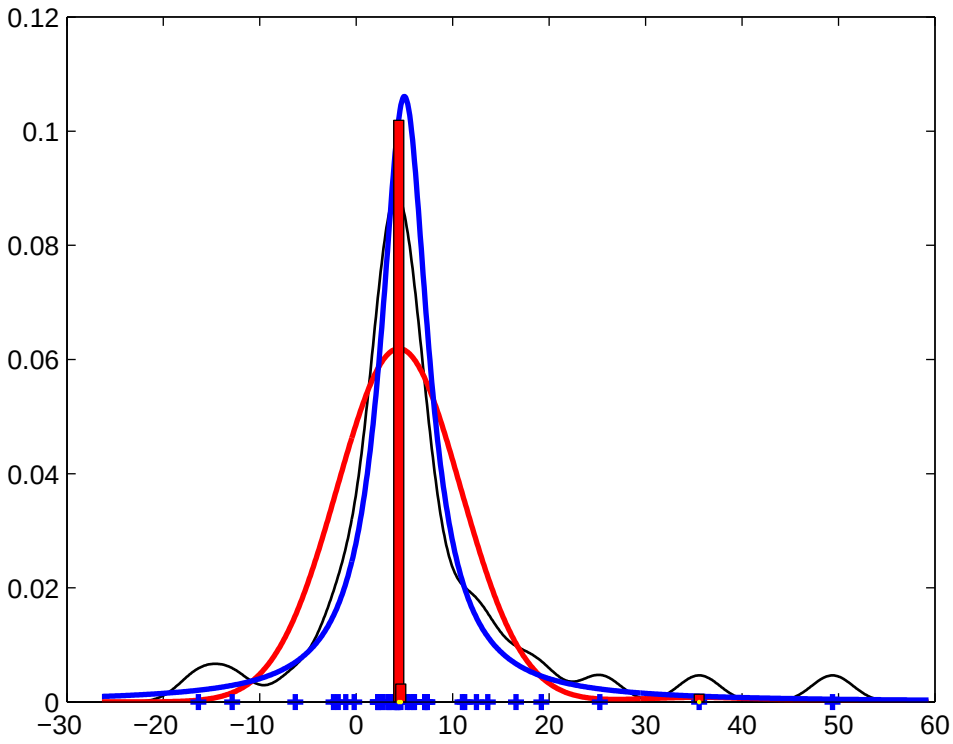
Example 17 (Heavy-tailed mixture) In this example 160 points from a mixture of Cauchy densities is considered.



Again the blue curve is the 'true' model from which the data was generated, the black curve is the Sheather-Jones estimator and the red curve is the GCE estimator. This time out the 160 point only 10 have a non-zero Lagrange multiplier. It is interesting to note that for heavy-tailed data the choice of the kernel function is significant. If, instead of a Gaussian kernel, we use a Cauchy kernel $K(x; X_i, \sigma) = \frac{1}{\pi\sigma}\frac{1}{1+(x-X_i)^2/\sigma^2}$ to estimate density, we get an almost perfect fit to the true Cauchy mixture (see next graph):

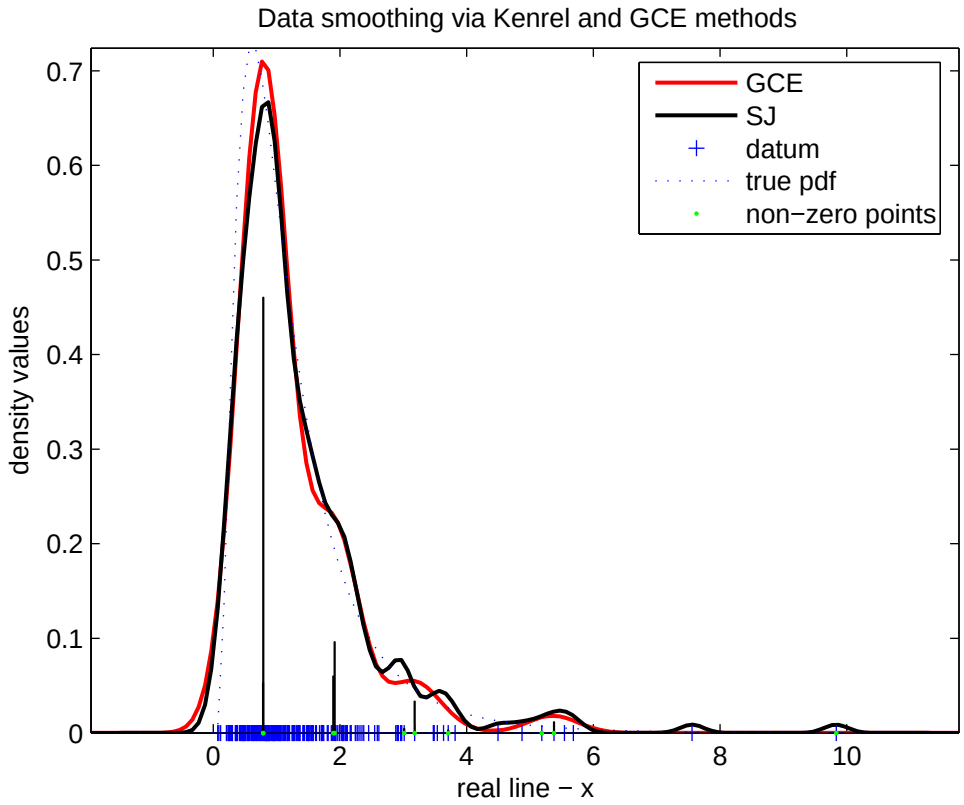


Example 18 (Robustness to Outlyers) In this example 40 points from a standard Cauchy density were generated. Only 3 points have non-zero Lagrange multipliers making the GCE estimator a Gaussian mixture with 3 components.

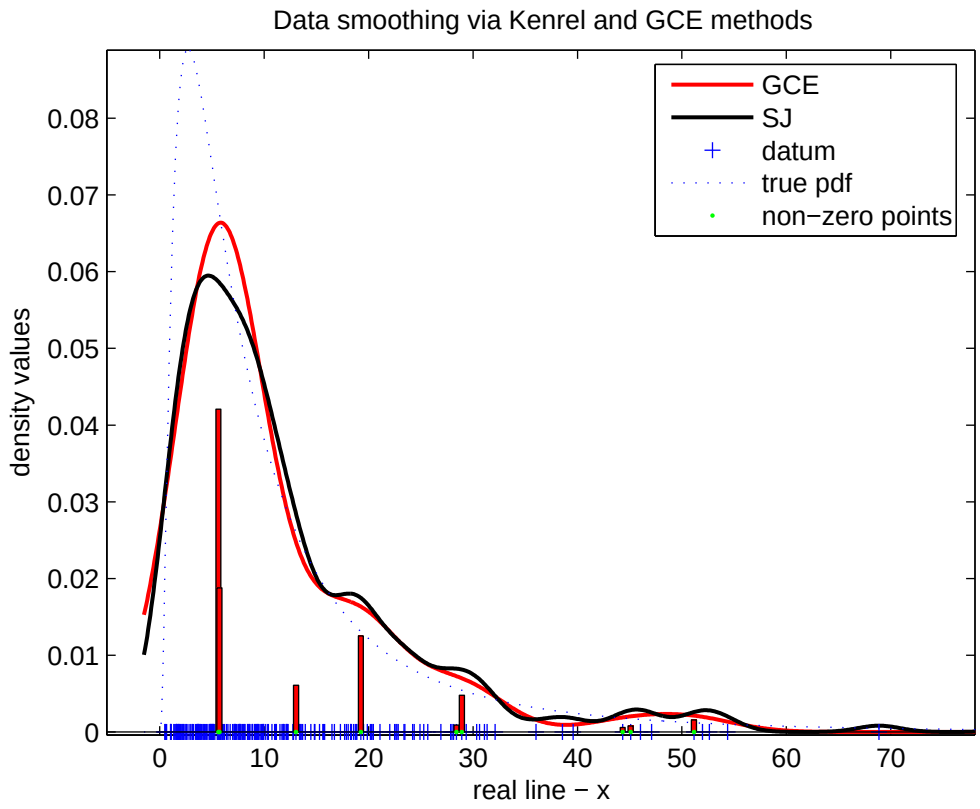


Note the spurious bumps in tails of the Sheather-Jones estimator. In general the GCE estimator is not sensitive to outlyers.

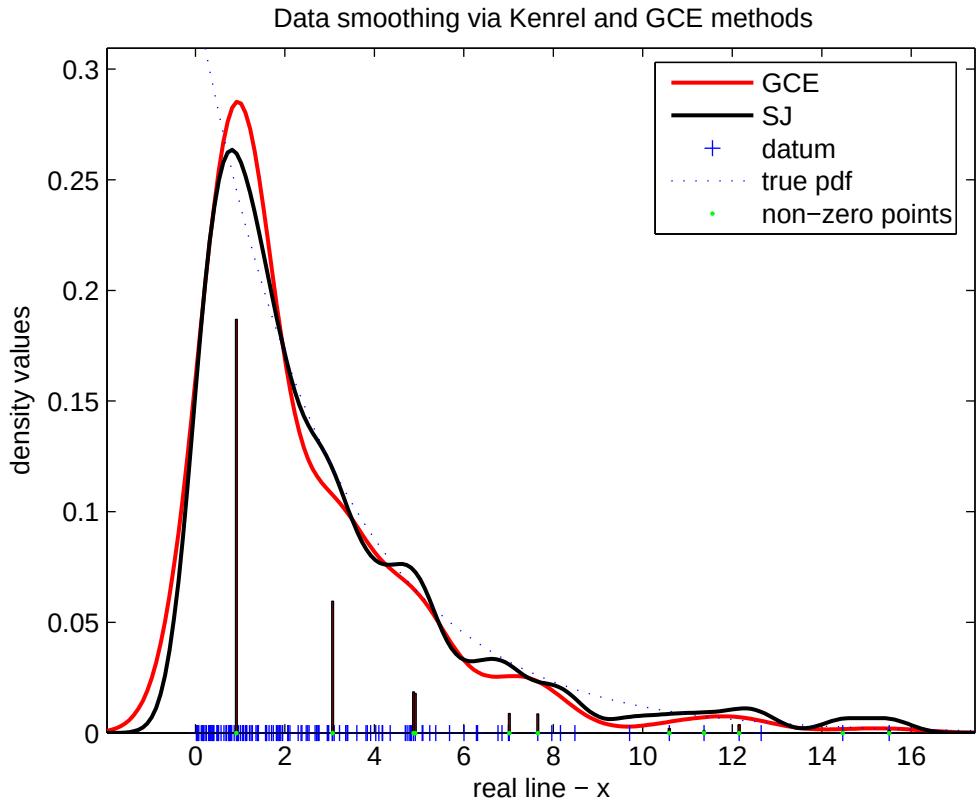
Example 19 (Lognormal Density) The first picture shows 240 points from the Lognormal density with location=0 and scale=.7. The green dots and the red bars show the data points associated with a non-zero Lagrange multiplier.



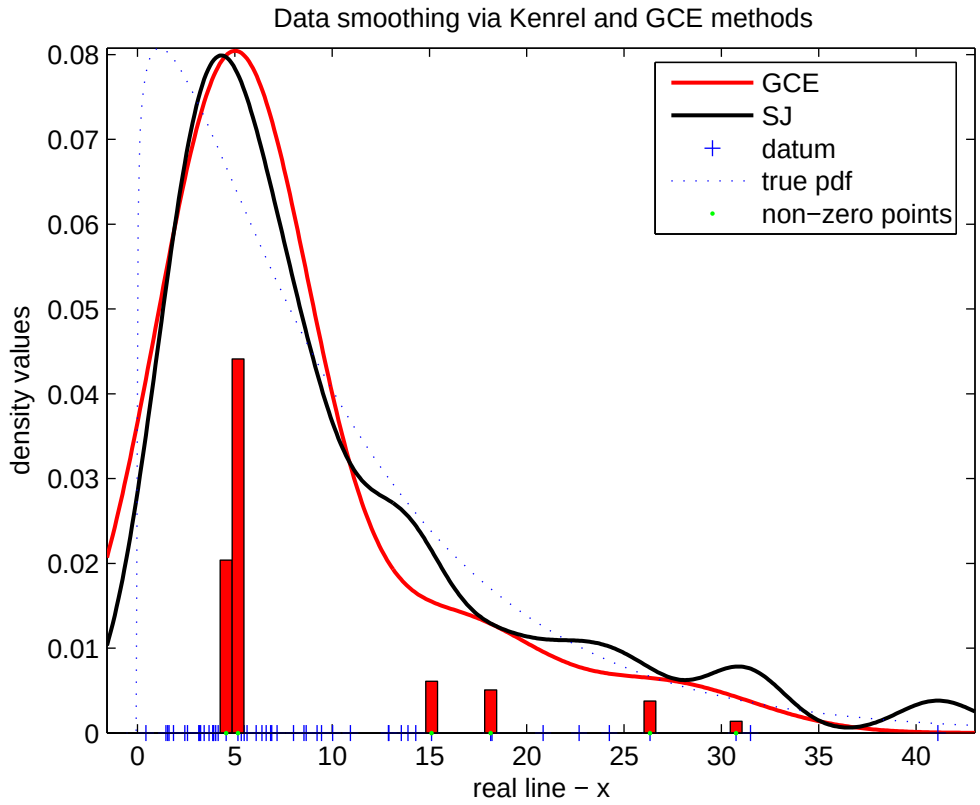
The second picture has 240 lognormally distributed points with scale=1 and location=2.



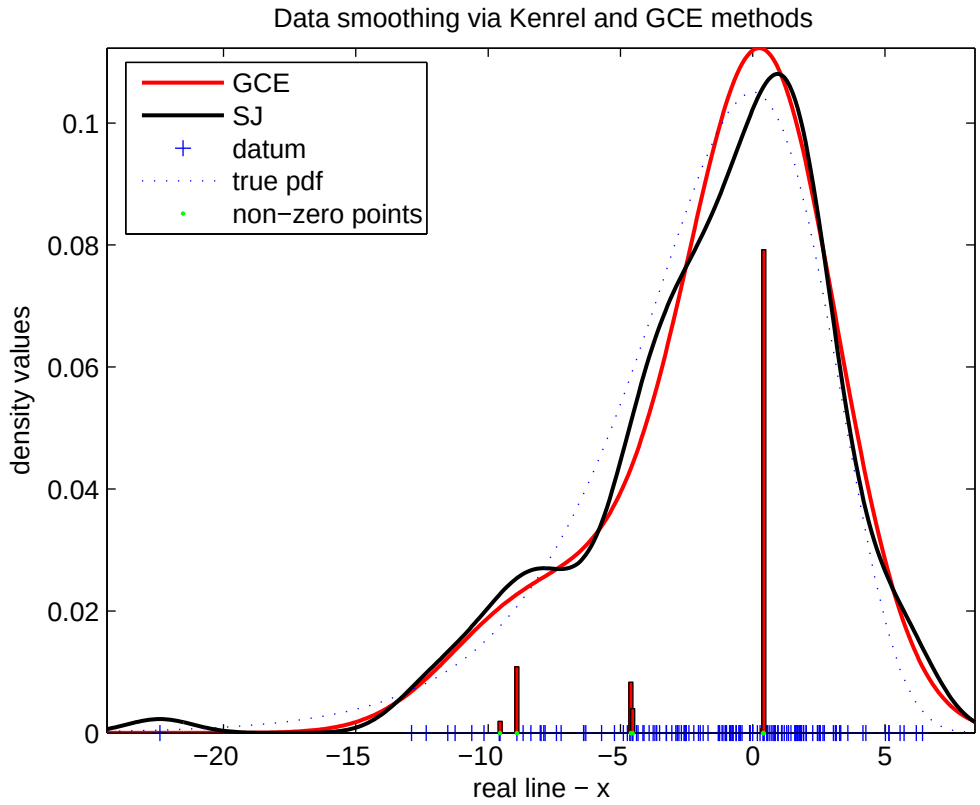
Example 20 (Weibull Density) The first picture shows 140 points from a Weibull density with location=3 and scale=1.



The second picture shows 50 points from a Weibull density with location=10 and scale=1.1 .



Example 21 (Extreme Value) This last example shows 140 points from an extreme value distribution with location=0 and scale=3.5 .



It is difficult to assess which of the two estimators is better. Both give reasonable and acceptable results. No attempt has been made to compare the two methods of density estimation beyond a visual subjective inspection. It is difficult to come up with a suitable measure of performance which will be fair to both methods. In conclusion:

1. The Sheather Jones estimator relies on the availability of large samples and essentially solves an asymptotic approximation approximately.

2. The derivation of the asymptotic approximations to MISE is valid only under the assumption that $X_1, \ldots, X_n$ are statistically independent, i.e., under the assumption that $\mathcal{X}_n \overset{i.i.d}{\sim} q^*$. The extension to the case of dependent observations is still an unsettled issue in the literature on kernel estimation (e.g., see [47], [24], [39], [40], [13], [12]). The GCE approach, however, does not make any assumptions about the statistical independence of the data.

3. The GCE solves the problem directly without using any asymptotic approximations[14]. The only approximation is in the estimation of the characterizing moments $\mathbb{E}_{q^*}[K_i(\mathbf{X})]$ through $\boldsymbol{\kappa}^*$. Apart from this approximation,

[14]The only other kernel method which does not rely on asymptotic theory is the Least

the GCE solves a functional optimization problem exactly to find the optimal density function.

4. The GCE gives a sparse mixture model.

5. Both methods attack the ill-posed problem of density estimation by introducing some external information in the probabilistic system. E.g., the Sheather-Jones method assumes differentiability of the unknown density q^* and independence of the data. The best method will ultimately be the one which imputes as little external information as possible to make the problem well-posed and provide a unique, stable and well-behaved density estimator.

Finally note that the use of the weighted kernel mixture (153) in density estimation is not novel. Hall & Turlach [23] first proposed the use of weights in density estimation and have successfully applied density estimators of the form (153). Later Girolami & He [21], [22] have applied the same idea to other statistical problems.

7 Discussion and Future Research

The original motivation for the GCE method is to solve difficult optimization and simulation problems. As an iterative learning algorithm a possible advantage of the GCE approach is that, unlike the CE method, the updating rules are fully automatic and the same for any of problems described in the introduction. The algorithm is like a black box—the only external information used is either random variables with distribution q^* or function values of q^* (up to a normalizing constant). It can, however, be also applied to standard statistical learning problems such as the problem of density estimation. The results of the numerical experiments are promising and show that the GCE method has the potential of becoming a powerful tool for tackling some of the most important problems in Statistics in a unified and simple framework. Some possible directions of future research are:

1. The Cross Entropy measures considered in this project can be derived axiomatically using basic concepts in Information Theory such as additivity and recursion. This will make the GCE method an axiomatically derived variational technique for solving ill-posed problems. It can then be argued that the GCE approach to statistical learning is as valid as the Bayesian approach. Bayesian statistical inference is also build axiomatically and in the absence of any inconsistencies there is no reason why we should prefer one set of axioms over another.

Squares Cross Validation method ([2], [63]) which unfortunately gives rough and spiky density estimates [68].

2. The GCE method seems to be related to the recently developed Support Vector Machines [66] and there is a possibility that the underlying principles of the Support Vector Machines can be derived via an information-theoretic approach.

3. A distinguishing feature of the GCE method is that it solves an infinite dimensional functional optimization problem. What makes this possible is the convexity and simplicity of the CE measures and the ensuing duality theory which allows us to reduce the variational problem to a finite parameter optimization problem. These results suggest that it may be possible to apply other more powerful Calculus of Variations techniques (or even Optimal Control) to the problems of Statistical Learning and Monte Carlo Simulation.

4. As a nonparametric density estimation methods the GCE provides an optimal non-asymptotic estimator. The consistency properties and the corresponding convergence rates of the GCE estimator need to be investigated.

5. All of the problems listed in the introductory section can be solved by (approximately) sampling from a suitably defined IS density function q^* via a CE-type algorithm [54] with iteratively updated levels. The GCE method has to be applied extensively on the problems of Monte Carlo Simulation and Statistical Learning.

References

[1] C. Lemarechal A. Decarreau, D. Hilhorst and J. Navaza. Dual methods in entropy maximization. applications to some problems in crystalography. *SIAM Journal of Optimization*, 1992.

[2] P. Hall A. W. Bowman and D. M. Titterington. Cross-validation in nonparametric estimation of probabilities and probability desnities. *Biometrika*, 71:341–351, 1984.

[3] I.S. Abramson. On bandwidth variation in kernel estimates—a square root law. *Ann. Stat.*, 10:1217–1223, 1982.

[4] J. Aczel. Measuring information beyond communication theory. *Inf. Proc. and Management*, 20:383–393, 1984.

[5] J. Aitchison and C.G.G. Aitken. Multivariate binary discrimination by the kernel method. *Biometrika*, 63:413–420, 1976.

[6] C. Arndt. *Information Measures: Information and its Description in Science and Engineering*. Springer, Germany, 2004.

[7] J. M. Borwein and A. S. Lewis. Duality relationships for entropy-like minimization problems. *SIAM J. Control and Optimization*, 29:325–338, 1991.

[8] J. M. Borwein and A. S. Lewis. *Convex analysis and Nonlinear Optimization: Theory and Examples*. Springer-Verlag, 2000.

[9] A. W. Bowman. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71:353–360, 1984.

[10] A. W. Bowman. A comparative study of some kernel-based nonparametric density estimators. *J. Statist. Comput. Simul.*, 21:313–327, 1985.

[11] J. P. Burg. The relationship between maximum entropy spectra and maximum likelihood spectra. *Modern Spectral Analysis*, 3:130–131,M.S.A., 1972.

[12] J. V. Castellana. Integrated consistency of smoothed probability density estimators for stationary sequences. *Stochastic Process. Appl.*, 33:335–346, 1989.

[13] J. V. Castellana and M. R. Leadbetter. On smoothed probability density estimation for stationary processes. *Stochastic Process. Appl.*, 21:179–193, 1986.

[14] S. T. Chiu. Bandwidth selection for kernel density estimation. *The Annals of Statistics*, 1991.

[15] T. F. Coleman and Y. Li. A reflective newton method for minimizing a quadratic function subject to bounds on some of the variables. *SIAM Journal of Optimization*, 6(4):1040–1058, 1996.

[16] I. Csiszár. A class of measures of informativity of observation channels. *Periodic Math. Hungarica*, 2:191–213, 1972.

[17] L. Devroye and L. Gyofri. *Nonparametric Density Estimation: The L_1 View*. Wiley Series In Probability And Mathematical Statistics, 1985.

[18] A. Doucet, N. de Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, New York, 2001.

[19] B. Efron and R. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1994.

[20] R. Fletcher. *Practical Methods of Optimization*. John Wiley and Sons, 1987.

[21] M. Girolami and C. He. Probability density estimation from optimally condensed data samples. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1253–1264, 2003.

[22] M. Girolami and C. He. Novelty detection employing an l_2 optimal non-parametric density estimator. *Pattern Recognition Letters*, 25:1389–1397, 2004.

[23] P. Hall and B. A. Turlach. Reducing bias in curve estimation by use of weights. *Computational Statistics and Data Analysis*, 30:67–86, 1999.

[24] J. D. Hart and P. Vieu. Data-driven bandwidth choice for density estimation based on dependent data. *Ann. Statist.*, 18:873–90, 1990.

[25] J. H. Havrda and F. Charvat. Quantification methods of classification processes:concepts of structural α entropy. *Kybernatica*, 3:30–35, 1967.

[26] E. T. Jaynes. Information theory and statistical mechanics. *Physical Reviews*, 106:621–630, 1957.

[27] J. N. Kapur. Measures of uncertainty, mathematical programming and physics. *Jour. Ind. Soc Ag. Stat.*, 24:47–66, 1972.

[28] J. N. Kapur. Four families of measures of entropy. *Indian Jour. of Pure and Applied Maths.*, 17:429–449, 1986.

[29] J. N. Kapur. New qualitative-quantitative measure of information. *Nat. Acad. Sci. Letters*, 6:51–54, 1986.

[30] J. N. Kapur. *Maximum Entropy Models in Science and Engineering*. Wiley Eastern, New Delhi, India, 1989.

[31] J. N. Kapur. Information-theoretic measures of stochastic dependence. *Bull. Math. Ind.*, 22:43–58, 1990.

[32] J. N. Kapur. *Measures of Information and Their Applications*. John Wiley & Sons, New Delhi, India, 1994.

[33] J. N. Kapur and H. K. Kesavan. *Generalized Maximum Entropy Principle (With applications)*. Standford Educational Press, University of Waterloo, Waterloo, Ontario, Canada, 1987.

[34] J. N. Kapur and H. K. Kesavan. The generalized maximum entropy principle. *IEEE Transactions on Syst., Man., and Cybernetics*, 19:1042–1052, 1989.

[35] J. N. Kapur and H. K. Kesavan. The inverse maxent and minxent principles and their applications. *P. F. Fougere (ed.), Maximum Entropy and Bayesian Methods*, Kluwer Academic Publishers:433–450, 1990.

[36] J. N. Kapur and H. K. Kesavan. Maximum entropy and minimum cross entropy principles: Need for a broader perspective. *P. F. Fougere (ed.), Maximum Entropy and Bayesian Methods*, Kluwer Academic Publishers:419–432, 1990.

[37] J. N. Kapur and H. K. Kesavan. On the family of solutions of generalized maximum and minimum cross-entropy models. *Int. J. Gen. Systems*, 16:199–219, 1990.

[38] J. N. Kapur and H. K. Kesavan. *Entropy Optimization Principles with Applications*. Academic Press, New York, 1992.

[39] T. Y. Kim. Asymptotically optimal bandwidth selection rules for the kernel density estimator with dependent observations. *J. Stat. Plann. Inf.*, 59:321–36, 1997.

[40] T. Y. Kim and D. D. Cox. A study on bandwidth selection in density estimation under dependence. *Journal of Multivariate Analysis*, 62:190–203, 1997.

[41] S. Kullback. *Information Theory and Statistics*. John Wiley & Sons, New York, 1959.

[42] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann.Math.Stat.*, 22:79–86, 1951.

[43] C. R. Loader. Bandwidth selection: Classical or plug-in. *The Annals of Statistics*, 27:415–438, 1999.

[44] J. S. Marron M. C. Jones and B. U. Park. A simple root n bandwidth selector. *The Annals of Statistics*, 19 4:1919–1932, 1991.

[45] J. S. Marron M. C. Jones and S. J. Sheather. Progress in data-based bandwidth selection for kernel density estimation. *Computational Statistics*, 11:337–381, 1996.

[46] J. S. Marron. An asymptotically efficient solution to the bandwidth problem of kernel density estimation. *The Annals of Statistics*, 13:1011–1023, 1985.

[47] S. N. Lahiri P. Hall and Y. K. Truong. On bandwidth choice for density estimation with dependent data. *Ann. Stat.*, 23:2241–63, 1995.

[48] L. A. Pars. *An Introduction to the Calculus of Variations*. Heinesmann. London, 1962.

[49] Y. Pawitan. *In All Likelihood: Statistical Modeling and Inference Using Likelihood*. Carendon Press Oxford, 2001.

[50] E. R. Pinch. *Optimal Control and the Calculus of Variations*. Oxford University Press, 1993.

[51] A. Renyi. On measures of entropy and information. *Proc. First Berkeley Symp. Stat.*, 1, 1961.

[52] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.*, 27, 1956.

[53] R. Y. Rubinstein. The stochastic minimum cross-entropy method for combinatorial optimization and rare-event estimation. *Methodology and Computing in Applied Probability*, 7:5–50, 2005.

[54] R. Y. Rubinstein and D. P. Kroese. *The Cross-Entropy Method*. Springer, 2004.

[55] M. Rudemo. Bias reduction in kernel density estimation by smoothed empirical transformations. *The Annals of Statistics*, 22:185–210, 1994.

[56] D. W. Scott. *Multivariate Density Estimation. Theory, Practice and Visualization*. John Wiley & Sons, 1992.

[57] A. K. Seth and J. N. Kapur. A comparative assessment of entropic and non-entropic methods of estimation. *P. F. Fougere (ed.), Maximum Entropy and Bayesian Methods*, Kluwer Academic Publishers:451–462, 1990.

[58] C. E. Shannon. A mathematical theory of communication. *Bell System Tech. J.*, 27:379–423;623–659, 1948.

[59] B. D. Sharma and D. P. Mittal. New non-additive measures of entropy for discrete probability distributions. *J. Math. Sci.*, 10:28–40, 1975.

[60] S. J. Sheather and M. C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *J.R.Statist.Soc.B*, 53:683–690, 1991.

[61] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, 1986.

[62] J. S. Simonoff. *Smoothing Methods in Statistics*. Springer, 1996.

[63] C. J. Stone. An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.*, 12, 1984.

[64] G. R. Terrell and David W. Scott. Variable kernel density estimation. *The Annals of Statistics*, 20:1236–1265, 1992.

[65] D.M. Titterington. A comparative study of kernel-based density estimates for categorical data. *Technometrics*, 22:259–268, 1980.

[66] V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.

[67] Frederick Y.M. Wan. *Introduction To The Calculus of Variations and Its Applications*. Chapman and Hall, 1995.

[68] M. P. Wand and M. C. Jones. *Kernel Smoothing*. Chapman & Hall, 1995.